

THE DIGITAL MUSE: ADVANCING LLM-BASED METHODS FOR
POETRY GENERATION AND AUTOMATED EVALUATION

A THESIS SUBMITTED TO
THE UNIVERSITY OF KENT
IN THE SUBJECT OF COMPUTER SCIENCE
THE SCHOOL OF COMPUTING
FOR THE DEGREE
OF PHD.

By
Piotr Sawicki
December 2025

Word count: 62541
Number of pages: 256

Abstract

This thesis investigates the application of Large Language Models (LLMs) to poetry generation and evaluation, chronicling methodological advancements during a period of unprecedented technological development in artificial intelligence (2021-2025). Through a series of interconnected studies spanning multiple generations of language models—from GPT-2 to Claude-3-Opus and GPT-4o—we develop frameworks for style-controlled poetry generation and automated evaluation that document both specific technical implementations and durable conceptual contributions.

Our research begins by examining the challenges of fine-tuning GPT-2 models to generate poetry in the style of specific Romantic-era poets, highlighting the importance of guarding against memorization and developing multi-faceted evaluation approaches. We then extend these methodologies to GPT-3, demonstrating the effectiveness of structured prompt-completion pairs for generating poetry with controlled content while preserving stylistic elements. Subsequent investigations reveal the limitations of zero-shot and many-shot prompting with early GPT-3.5-turbo and GPT-4 models, emphasizing the continued importance of fine-tuning for specialized stylistic tasks at that technological stage.

In later chapters, we shift focus to the evaluation challenge, developing a methodology inspired by the Consensual Assessment Technique (CAT) that leverages state-of-the-art LLMs as judges of poetic quality. We demonstrate that these models can significantly outperform non-expert human judges in aligning with established ground truth quality rankings, providing a reliable and scalable alternative to traditional evaluation methods. Finally, we apply these generation and evaluation methodologies to compare the quality of AI-generated and human-written poetry, finding that, according to LLM evaluators, recent AI models can produce poems matching or exceeding certain categories of human poetry. However, this assessment reveals

significant differences in evaluation biases between LLMs and underscores the need for further validation by human literary experts.

The primary contributions of this thesis include:

1. A demonstrably effective methodology for fine-tuning LLMs with structured summary-poem pairs to generate user-controlled, stylistically consistent poetry.
2. A novel LLM-based, CAT-inspired in-context evaluation framework for assessing individual poems across multiple criteria.
3. Empirical evidence that LLM-based evaluators can surpass non-expert human judges in aligning with poetry quality benchmarks when assessing human-written works.
4. Systematic documentation and analysis of the evolving capabilities of successive LLM generations in creative text generation and evaluation.
5. Identification and critical analysis of significant evaluation biases between state-of-the-art LLMs (Claude-3-Opus and GPT-4o). This analysis reveals that while LLMs can reliably tier human poetry, their assessment of AI-generated poetry may reflect preferences for 'LLM-native' characteristics, underscoring that high LLM-assigned quality scores for AI work necessitate careful interpretation and validation by human literary experts.

Collectively, these contributions advance the state of the art in computational poetry generation while establishing methodological frameworks that maintain relevance despite the rapid evolution of underlying language model technologies.

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Marek Grześ and Professor Dan Brown, for their continuous support throughout my doctoral studies. They have been dedicated guides during this entire journey, always generous with their time and advice. I am particularly grateful for the countless hours they spent assisting me and for the collaborative work that resulted in our joint papers. I also extend my sincere thanks to Dr. Anna Jordanous. Although not officially my supervisor, she helped immensely over the course of my PhD, acting effectively as a third supervisor.

I would like to acknowledge the collaboration with Dr. Fabrício Góes from the University of Leicester; we shared many meaningful conversations and contributed to each other's research. Similarly, I thank Max Peepkorn and Aisha Khatun for our collaboration and insightful discussions.

My PhD journey began thanks to the guidance of several staff members at the University of Kent. I am grateful to Dr. Laura Bocchi, Professor Andy King, Professor Colin Johnson, and Professor Peter Rodgers for those initial pivotal conversations that helped me get started on this path.

I also wish to acknowledge the support of Professor Alex Freitas and Dr. Rogério de Lemos, who served on my yearly PhD review panel. Their helpful corrections and guidance throughout the process were invaluable.

My thanks also go to Sonnary Dearden and the entire administrative team for their unwavering support over the years, as well as to Tim Bishop, our systems administrator, for his expert technical assistance whenever I needed it. The School of Computing at the University of Kent has been a wonderful environment in which to undertake a PhD, fostering a research atmosphere

that is simply fantastic.

Finally, I would like to thank my external examiners, Dr. Hugo Gonçalo Oliveira and Dr. Carlos León Aznar, for their constructive feedback, which greatly improved the quality of this thesis. I also thank my internal examiner, Professor Peter Rodgers, for his continued support and meaningful advice.

Contents

Abstract	ii
Acknowledgements	iv
Contents	vi
List of Tables	xii
List of Figures	xvi
1 Introduction	1
1.1 Research Focus	1
1.2 Motivation: Poetry as a Frontier for Controllable Creative AI	2
1.3 Operationalizing “Style” in Computational Generation	4
1.4 Original Contributions	5
1.5 Evolution of Generated Outputs: From Incoherent Fragments to Full Compositions	6
1.6 Structure of the Thesis	11
1.7 Publications Derived from this Research	12
1.8 Research Context and Trajectory: Navigating the Rapidly Evolving LLM Land- scape	13
1.9 Conclusion	15
2 Background	16
2.1 General Introduction to Large Language Models	16
2.1.1 Introduction to Word Embeddings	17

2.1.2	Static Word Embeddings	17
2.1.3	Dynamic Word Embeddings	18
2.2	Large Language Models	19
2.2.1	Sequential Models: RNNs, LSTMs, and the Challenge of Context	19
2.2.2	The Sequence-to-Sequence Framework and its Bottleneck	20
2.2.3	Attention Mechanisms	20
2.2.4	The Transformer Architecture	22
2.2.5	GPT Models	25
2.2.6	BERT and Its Variants	28
2.2.7	Adapting Pre-trained Language Models: Fine-Tuning and Prompting . .	30
2.2.8	ChatGPT and onwards	31
2.2.9	Open-source LLMs	33
2.3	Theoretical Perspectives and Evaluation Frameworks for Computational Creativity	34
2.3.1	The Question of Intentionality in Creative Agency	35
2.3.2	Defining Creativity	37
2.3.3	Creativity Evaluation Frameworks	39
2.3.4	Creativity in Computational Poetry Generation	42
2.3.5	The Role of Value and Intention: Mere Generation?	43
2.3.6	Evaluation in Practice: Challenges and Approaches	45
2.3.7	Implications for Computational Poetry Generation	46
2.3.8	Conclusion: Theoretical Foundations for Computational Poetry	47
2.3.9	Linking Creativity Theory to Thesis Methodology	48
2.4	The Evolution of Digital Poetry Generation in the Pre-LLM Era	49
2.4.1	Early Approaches: Rule-Based Systems and Traditional NLP Methods .	50
2.4.2	Frameworks for Understanding Computational Poetry Generation	51
2.4.3	Evolution Towards Sophisticated Approaches	51
2.4.4	Evaluating Computational Poetry	56
2.4.5	Interactive Systems and Creative Support	56
2.4.6	Conclusion: Evolution of Digital Poetry Generation	57
2.5	Digital Poetry Generation in the LLM Era	57

2.5.1	Introduction	57
2.5.2	Foundational Models and Approaches	58
2.5.3	Addressing Specific Poetic Constraints and Forms	59
2.5.4	Advanced Techniques and Architectures	60
2.5.5	Human-AI Collaboration	60
2.6	Evaluation Methods	61
2.6.1	Automated Evaluation	61
2.6.2	Human Ability to Differentiate Between AI and Human Poetry	61
2.6.3	Depth of Creativity and Emotional Resonance	62
2.6.4	Bias and Ethical Issues	63
2.7	Conclusion	63
3	Training GPT-2 for Romantic-Era Poetry Generation	65
3.1	Defining Authorial Style in Computational Generation	66
3.2	GPT-2 Models and Experimental Setup	66
3.3	Data preparation	68
3.3.1	Original datasets	68
3.3.2	Setup 1 for visual and BLEU evaluations	69
3.3.3	Setup 2 for BERT evaluations	69
3.4	Part 1—Evaluation of the overfitted models	71
3.4.1	Qualitative Overview of Generated Output	71
3.4.2	Experiment 1—Visual evaluation of text quality	78
3.4.3	Experiment 2—What does the BLEU evaluation tell us?	82
3.5	Part 2—BERT evaluation of produced samples	88
3.5.1	Experiment 3—Can fine-tuned GPT-2 outwit BERT?	89
3.5.2	Experiment 4—Which GPT-2 is better at replicating the author?	91
3.6	Discussion	93
3.7	Conclusion	94
4	Fine-Tuning GPT-3 For Poetic Style Generation	96
4.1	Introduction	96

4.2	Objectives and Methods	99
4.2.1	Research Objectives	99
4.2.2	Evolution of Methodology: Technological Enablers and Design Shifts	99
4.2.3	Contributions and Workflow	100
4.3	Part 1—Poetry Generation	101
4.3.1	Data Preparation	102
4.3.2	Fine-tuning GPT-3 for Poetry Generation	104
4.3.3	Generating Poems from the Fine-tuned GTP-3 Models	106
4.4	Part 2—Poetry Evaluation	107
4.4.1	Step 1—Establishing the Accuracy of GPT-3-based Classifiers	109
4.4.2	Step 2—Evaluating GPT-3-generated Poetry Against the Works of the Original Author Using GPT-3-based Classifiers	113
4.4.3	Qualitative Overview of Generated Poems	116
4.5	Discussion	128
4.6	Conclusion	129
5	Prompting for Stylistic Poetry Generation	131
5.1	Method	132
5.1.1	Three Different Models and One Prompt	132
5.1.2	Do GPT models Know Whitman’s Poems?	133
5.1.3	Experimental Setup	136
5.1.4	Data Preparation	139
5.1.5	Evaluation	141
5.2	Discussion and Conclusion	142
6	LLM-Based Evaluation Of Poetry	144
6.1	Methodology	145
6.2	Dataset	149
6.3	Results of Human Evaluation	150
6.4	Experiment 1—Evaluating Poems Without Context	150
6.5	In-context Poem Evaluation	153

6.5.1	Analyzing Input Size for In-context Poem Evaluation	154
6.5.2	Experiment 2—Evaluation of Poems—90 poems in a prompt	155
6.5.3	Experiment 3—Evaluation of Poems—15 poems in a prompt	155
6.6	Experiment 4—Evaluating Poetry - Interrater Reliability of Claude-3-Opus and GPT-4o	159
6.7	Comparison with Non-Expert Human Evaluations	159
6.8	Further Discussion and Implications	160
6.9	Conclusion	161
7	Human Poetry vs AI Poetry: LLM Evaluation	166
7.1	Datasets	168
7.1.1	Fine-tuning dataset	168
7.1.2	Test Summaries for Poem Generation	169
7.2	Methods for Generating Poems	170
7.2.1	Generating Poems Through Prompts Only	170
7.2.2	Generating Poems with Fine-Tuned GPT-4o Models	171
7.2.3	Refining AI-generated Poems from Fine-Tuned Models	171
7.3	Selection of AI-Generated Poems for Evaluation	172
7.3.1	Addressing Plagiarized Content in Selected Fine-Tuned Poems	175
7.3.2	Best poems from each of the models	175
7.4	Comparing AI-generated Poems to Human Poetry	179
7.4.1	Example poems from evaluations of AI-generated poems against human poetry	184
7.5	Comparing LLM-generated Poems Across Models	194
7.5.1	Evaluating Four AI Generation Methods	194
7.5.2	Highest scoring poems from evaluating four datasets of AI-generated poetry	197
7.5.3	Evaluating Top vs. Bottom Poems Across Methods	199
7.5.4	Poems from the Top vs. Bottom Quality Comparison	203
7.6	Inter-rater Reliability of LLM Quality Evaluations	212
7.7	Discussion and Conclusion	213

8	Conclusions	218
8.1	Key Contributions within an Evolving Methodological Landscape	218
8.2	Directions for Future Research	220
8.3	Final Reflections	221
	Bibliography	223

List of Tables

1	Statistical summary of the datasets used for fine-tuning. The table compares the original collected works (Org.) against the final pre-processed versions, detailing file sizes (in MB) and total line counts.	68
2	Results of Experiment 4: BERT classification of original poet’s work using classifiers trained on pairs of GPT-2 generated data. The winning model (closer to the original author’s style) is indicated in bold based on the accuracy score. .	92
3	Fine-tuning configurations for models trained on individual poets (300 poems each, top table) and combined poets (7 poets, 2100 poems, indicated by “7A”, bottom table). The “Dataset and Epochs Code” notation is used throughout the chapter to identify specific model configurations.	106
4	Example prompt used to generate the poem presented in Table 5.	107
5	Example poem generated by our system using the prompt shown in Table 4. . .	108
6	Results of evaluating the accuracy of GPT-3-based binary classifiers in Step 1 (Section 4.4.1).	110
7	Classification accuracy in distinguishing GPT-3-generated poetry from original works. Models with the best style preservation performance (lowest accuracy scores) for each poet are highlighted in bold. The model codes in the first column indicate the base model, dataset type (7A = seven-author dataset), and fine-tuning epochs (1e/4e).	114

8 Example of the first stanzas of the poems generated by GPT-3.5, GPT-3.5-turbo and GPT-4 from the prompt shown, followed by the fragment of a poem generated from the same summary from the fine-tuned GPT-3 Curie. The summary provided in the prompt is of the poem ‘Out Of The Night That Covers Me’ by William Ernest Henley. For the fine-tuned model, the prompt did not include the verbal instruction. 134

9 Results of retrieving the complete text of the poems by our chosen author. The values represent similarity scores derived from Levenshtein distance, calculated over five trials. Higher percentages indicate greater similarity between the retrieved text and the original poems. A score of 100% would indicate perfect retrieval, while 0% would indicate complete dissimilarity. 135

10 Comparison between original Whitman poem “Who Learns My Lesson Complete?” and GPT-3.5 generated version with similarity analysis 137

11 Comparison between original Whitman poem “Spirit Whose Work Is Done” and GPT-4 generated version with similarity analysis. Both poems have been truncated to fit the page. The analysis results are based on the complete texts. . . 138

12 GPT versions used for poetry generation. 139

13 Results of our experiments where GPT-generated poetry is compared against the Walt Whitman’s original works. Entries in the first column indicate which GPT model’s output was evaluated against the Whitman’s works. 142

14 Spearman’s Rank Correlation (SRC) and p-values from human non-expert judges evaluations (Lamb, Brown and Clarke 2015), ordered by SRC descending. . . . 151

15 Descriptions of the selected evaluation criteria by non-expert human judges presented in (Lamb, Brown and Clarke 2015) 151

16 Results of Experiment 1 for all models. The number of poems that were categorized to the given category (correct predictions are in brackets). 152

17 Results of Spearman’s Rank Correlation (SRC) against the ground truth of the poems ordering in 5 criteria from 90-poems prompts using Claude-3-Opus (Experiment 2, “Claude-3-Opus 90n” and “Claude-3-Opus 90”) and 15-poems prompts (Experiment 3, “Claude-3-Opus 15”, “Claude-3-Opus”, “GPT-4o 15n” and “GPT-4o”). The best result for each model is presented in bold. For comparison with human evaluations please see Table 14. 154

18 ANOVA results for poem evaluation in five criteria in Experiments 2 and 3 for all models with both scoring methods. Columns “Good,” “Medium,” and “Bad” present averaged scores for those poems’ categories. 158

19 Example of 10 evaluations of the same subset of 15 poems in Experiment 4 with Claude-3-Opus 15n and Claude-3-Opus, evaluating the criterion “Creativity”. Poem’s numbers indicate the categories they belong to: 1-30 “Good” (A), 31-60 “Medium” (B), 61-90 “Bad” (C). These results confirm that while the temperature parameter successfully diversifies the output of its LMM, the ICC results Table. 20 show that those diversified outputs are consistent and reliable on average. 164

20 Results of ICC evaluations in Experiment 4, computed for both models with both scoring approaches and for the five criteria. 165

21 Demographic information and poem counts by author for fine-tuning and test summary datasets. The “Poems in Fine-tuning Set” column indicates the number of poems by each author included in the 2000-poem fine-tuning corpus. The “Poems in Test Summary Set” column indicates the number of poems by each author among the 100 poems reserved for generating test summaries. 168

22 Comparison between a raw text of a poem “From Golden Walls to Rooted Prayers” generated by FT-GPT-4-mini (based on the provided summary) and its refined version by GPT-4o. 173

23 Comparison between a raw text of a poem “Darkening Air Deeper Delight” generated by FT-GPT-4o (based on the provided summary) and its refined version. 174

24 Plagiarized poems identified in the initial top selection from fine-tuned models . 176

25	ANOVA Results: Comparison of Mean Quality Scores between Human Poem Categories (A, B, C) and LLM-Generated Poems (D) for Section 7.4 Experiments. Scores are Mean \pm STD.	180
26	ANOVA Results: Comparison of Mean Quality Scores among Four AI Generation Methods (Section 7.5.1). Each row represents a separate ANOVA comparing the four AI methods listed as columns, evaluated by the specified LLM and using either refined or raw versions of fine-tuned poems. Degrees of freedom for F-statistic are (3, 116) for all tests. The “Ref.” or “Raw” in the “Evaluation (Condition)” column indicates whether the refined or raw versions of the fine-tuned poems were included in that specific 4-way comparison.	195
27	T-test Results: Comparison of Mean Quality Scores between pre-selected “Top” (N=8) and “Bottom” (N=8) AI-Generated Poems (Section 7.3). “M Top (SD)” refers to Mean (Standard Deviation) for the “Top” group; “M Bottom (SD)” for the “Bottom” group. The “(Raw FT versions)” or “(Refined FT versions)” indicates the status of fine-tuned poems in the set. Degrees of freedom for t-tests are \approx 14 (Welch’s t-test).	200
28	LLM Poem Scores with updated data (raw versions from FT models). <i>Top-ranked entries are shown in bold, bottom-ranked entries in italics.</i>	201
29	LLM Poem Scores for top and bottom poems (refined versions from FT models). <i>Top-ranked entries are shown in bold, bottom-ranked entries in italics.</i>	202
30	ICC evaluation results across all Chapter 7 experiments (average rater reliability measures)	214

List of Figures

1	Overview of the Sequence-to-Sequence (Seq2Seq) or Encoder-Decoder model. .	20
2	Overview of the Sequence-to-Sequence (Seq2Seq) or Encoder-Decoder model with Attention mechanism.	21
3	Attention mechanism - scoring	22
4	Attention mechanism - context vector.	22
5	Final step of Attention mechanism.	23
6	Transformer model architecture (Vaswani et al. 2017)	24
7	Transformer architecture adapted to various NLP tasks (Radford et al. 2018) . .	27
8	Adapting the pre-trained GPT Transformer architecture for various NLP tasks during fine-tuning. Minimal task-specific layers are added to leverage the model’s learned representations for classification, entailment, similarity, and multiple-choice problems (Radford et al. 2018).	28
9	BERT training objectives (Devlin et al. 2018)	29
10	Training process of ChatGPT (OpenAI 2022).	32
11	Training and evaluation loss for the LMH Small model (Top) and the Regular Small model (Bottom) fine-tuned for 250K steps on the dataset of Collected Works of Lord Byron.	70
12	Results of the visual evaluation of text quality for samples generated from the Byron (left) and Shelley (right) datasets.	79
13	BLEU scores for Byron (left) and Shelley (right) calculated for samples with 1000 tokens length.	83

14	Results of the BERT evaluation of the Byron (left) and Shelley (right) checkpoint samples produced by the four different GPT-2 models. Row 1 (top): Regular Medium models, Row 2: Regular Small, Row 3: LMH Medium, Row 4 (bottom): LMH Small.	90
15	Combined methodological workflow. The left “Generation” panel details the process of fine-tuning an LLM (GPT-3) with summary-poem pairs to generate new poems in an original author’s style. The right “Evaluation” panel illustrates how these generated poems (as Label 0) are then compared against the original author’s poems (Label 1) by a fine-tuned LLM classifier to assess stylistic fidelity, aiming for an accuracy score of 50%	98
16	Raw corrupted input text for Sample #44, misclassified by the Ada model in the Walt Whitman vs. Machine Learning Book task.	112
17	Raw corrupted input text for Sample #141, misclassified by the Davinci model in the Walt Whitman vs. Machine Learning Book task.	112
18	The poem orders obtained in Experiments 2 and 3 on all criteria, where every poem is colour coded with its ground truth category. The lower part of the table presents for comparison the orderings by human non-expert judges from Lamb, Brown and Clarke (2015).	156
19	Ordering of the LLM-generated poems vs. human poems of three categories. LLM poetry generated by Claude and GPT-4o, evaluation done by Claude and GPT-4o.	181
20	Ordering of the LLM-generated poems by Claude, GPT-4o, FT-GPT-4o and FT-GPT-4o-mini.	195

Chapter 1

Introduction

1.1 Research Focus

This thesis delves into the challenge and potential of using Large Language Models (LLMs) for generating and evaluating high-quality poetry. While LLMs demonstrate remarkable capabilities across many Natural Language Processing (NLP) tasks, crafting poetry that is not only aesthetically compelling but also adheres to specific stylistic conventions or user-defined content remains a complex frontier. Addressing these complexities, this research pursues solutions to two primary research questions:

1. How can we develop LLM-based methods to generate poetry that maintains user-controlled content while faithfully adhering to the distinctive style of a specific poet?
2. How can we establish robust and reliable methodologies for the automated evaluation of poetic quality, applicable to both human-written and computer-generated works?

Answering these questions requires navigating a rapidly evolving technological landscape (2021-2025), where the capabilities of LLMs have advanced at an unprecedented pace. Therefore, a key part of this investigation involves not only developing novel generation and evaluation techniques but also critically examining their performance across different model generations, understanding their limitations, and critically examining the nuances inherent in using AI for subjective tasks like creative assessment. This includes investigating potential variations,

inherent biases, and the specific nature of quality judgments when LLMs evaluate both human and AI-generated creative works.

This introductory chapter sets the stage for this investigation. It will further detail the specific research focus, outline the original contributions stemming from this work, explain the chronological structure adopted for the thesis, list the publications derived from this research, and explicitly contextualize the unique challenges and opportunities presented by conducting research during this period of rapid AI development. The chapter concludes by establishing the methodological philosophy underpinning the thesis – one aimed at ensuring the durability of findings and insights despite the transient nature of specific technical implementations.

1.2 Motivation: Poetry as a Frontier for Controllable Creative AI

The intersection of Large Language Models (LLMs) and poetic composition represents more than a novel application of technology to art; it constitutes a critical frontier in the field of Computational Creativity. While poetry is often regarded as the most subjective and human of all literary forms—relying on ambiguity, emotion, and lived experience—LLMs are fundamentally probabilistic engines designed to predict the next token in a sequence.

The motivation for this research lies precisely in this tension between the statistical nature of the machine and the semantic density of the poem. By building systems capable of generating digital poetry, we are not merely automating an artistic practice; we are probing the limits of machine understanding and creating new frameworks for human-computer interaction. This thesis is driven by several key motivations, ranging from technical NLP challenges to societal transparency.

Poetry as a Test of Semantic Understanding

From an Artificial Intelligence perspective, poetry presents a unique adversarial challenge for NLP systems. Standard tasks such as summarization, translation, or information retrieval generally prioritize precision and factual disambiguation. Poetry, conversely, prizes ambiguity, metaphor, novel juxtaposition, and strict structural constraints. Therefore, building systems that generate poetry serves to stress-test the capabilities of LLMs beyond functional communication.

This is particularly relevant when addressing the phenomenon of “hallucination.” In a factual

context, an LLM generating false associations is considered a failure; however, in a poetic context, a novel or unexpected association is often a successful metaphor. Researching how to control this distinction improves our ability to steer LLMs toward creativity or accuracy as needed. Furthermore, the genre imposes complex constraint satisfaction problems—requiring the model to maintain semantic coherence while simultaneously adhering to meter, rhyme schemes, or visual layouts—which contributes to broader AI research regarding constrained generation and controllable text synthesis.

Interdisciplinary and Methodological Benefits

Beyond computer science, research into digital poetry contributes to the Digital Humanities and Media Studies by offering empirical data on how algorithmic processes influence narrative and aesthetic theory. It bridges the gap between the “Two Cultures” of science and the humanities by applying rigorous computational methods to aesthetic inquiry, allowing us to quantify and analyze the mechanics of creativity in ways previously unavailable to literary scholars.

Instrumental Augmentation: AI as an Assistive Tool

A central motivation of this thesis is to investigate the use of AI as a sophisticated assistive tool for the human writer, rather than as an independent creator. In this framework, the agency remains entirely with the human author, with the system functioning as a mechanism for expansion and support. Moreover, as an assistive scaffold, these systems have the potential to democratize expression. They allow individuals who may lack technical writing skills—due to language barriers or cognitive differences—to formulate complex emotional states by relying on the system for structural support.

Societal Implications: Algorithmic Interpretability and Cultural Heritage

Finally, the societal relevance of research into digital poetry extends to the transparency of algorithmic systems. As AI becomes ubiquitous, understanding how it processes language is vital. Poetry, by stripping away the veneer of factual utility, reveals the raw “thinking” process of the model. Analyzing generated poetry serves as a form of “algorithmic interpretability,” exposing the model’s biases, training data limitations, and latent associations. Additionally, because LLMs are trained on vast archives of human culture, generating poetry becomes a method of “remixing”

and re-contextualizing this heritage, potentially preserving and revitalizing older literary forms through digital regeneration.

In summary, the development of LLM-driven digital poetry is not a trivial pursuit of robotic entertainment. It is a rigorous inquiry into the boundaries of Natural Language Processing, a design challenge for future creative tools, and a method for investigating the statistical substrate of modern AI.

1.3 Operationalizing “Style” in Computational Generation

Central to this thesis is the specific definition of “style” as a computational objective. In the context of our experiments, we do not view style merely as adherence to a fixed set of poetic constraints—such as the rhyme scheme of a sonnet or the meter of a haiku—but rather as the distinctive linguistic signature of a specific author. It is the complex, often tacit interplay of lexical choices, syntactic habits, punctuation idiosyncrasies, and recurring thematic preoccupations that constitutes an author’s unique voice.

For a Large Language Model, replicating this style requires more than following explicit instructions; it demands a fundamental shift in the model’s probability distribution. We treat style as a learnable feature set where the objective is to predict the specific word choices and rhythmic cadences that a target author, such as Lord Byron or Walt Whitman, would employ in a given context. This operational definition moves beyond simple mimicry of form. It challenges the model to capture the texture of the author’s language—the archaic diction of the Romantics, the expansive free verse of the American transcendentalists, or the specific emotional resonance of their imagery.

By framing style in this way, we transform the subjective literary challenge of “writing like an author” into a rigorous, evaluable technical task. Success is not measured by whether the output looks like a poem, but by whether the generated text creates a statistical and semantic illusion compelling enough to be indistinguishable from the author’s genuine corpus, even when the content itself is novel. This distinction between the *container* of style and the *substance* of content forms the basis for the fine-tuning and evaluation methodologies developed throughout this research.

1.4 Original Contributions

This thesis presents several original contributions to the fields of computational creativity and natural language processing, particularly concerning LLM-based poetry generation and evaluation. Initial investigations highlighted critical evaluation challenges with early models, such as distinguishing genuine stylistic mimicry from plagiarism and the limitations of standard metrics (Chapter 3). This, combined with a subsequent empirical assessment of prompting capabilities (Chapter 5) which revealed their limitations for complex stylistic control at that technological stage, motivated the development of the core methodologies and evaluation frameworks presented below.

The primary original contributions of this research include:

1. A demonstrably effective methodology for fine-tuning LLMs using structured summary-poem pairs to generate poetry that adheres to both user-controlled content specifications and specific authorial styles (Chapter 4).
2. An empirical investigation into the efficacy of zero-shot and many-shot prompting with early advanced LLMs (GPT-3.5, GPT-4) for achieving specific authorial style replication in poetry, establishing their insufficiency for this specialized task at that stage and thereby underscoring the continued relevance of targeted fine-tuning approaches (Chapter 5).
3. A novel in-context evaluation framework, inspired by the Consensual Assessment Technique, for assessing individual poems across multiple criteria using LLMs as surrogate expert judges (Chapter 6).
4. Compelling empirical evidence, derived from the application of the evaluation framework, demonstrating that LLM-based evaluators can surpass non-expert human judges in poetry assessment reliability and alignment with established quality benchmarks when assessing human-written poetry (Chapter 6).
5. Systematic documentation and analysis of the evolving capabilities of successive LLM generations (from GPT-2 through Claude-3-Opus/GPT-4o) in the specialized domains of creative text generation and evaluation during the pivotal 2021-2025 period (contextualized in Chapters 1 & 2, with evolution demonstrated across Chapters 3-7).

6. Empirical identification and analysis of significant differences in evaluation biases between state-of-the-art LLMs (Claude-3-Opus and GPT-4o). This includes exploring how LLMs assess AI-generated versus human-written poetry, revealing potential LLM preferences for certain generative characteristics and underscoring the critical need for human expert validation in interpreting AI-driven quality assessments of creative works (Chapter 7).

Collectively, these contributions advance the state-of-the-art by providing practical methods for controlled creative generation, establishing robust and scalable automated evaluation techniques, and offering valuable insights into the rapidly progressing capabilities and limitations of LLMs in artistic domains.

1.5 Evolution of Generated Outputs: From Incoherent Fragments to Full Compositions

To fully appreciate the technical trajectory of this research, it is necessary to visualize the concrete evolution of the generated outputs. As the underlying models advanced from GPT-2 to GPT-3 and eventually GPT-4o, the capacity of the system shifted from generating stylistically coherent fragments to composing structurally complete and thematically controlled poems.

Phase 1: Stylistic Fragments (GPT-2). As detailed in **Chapter 3**, our initial work with GPT-2 focused on the texture of style rather than structural coherence. Due to the limitations of the model size and the training objectives available in 2021, the system did not reliably produce complete poems with a distinct beginning, middle, and end. Instead, it produced continuous streams of text mimicking the target author’s voice, from which stylistically representative fragments could be extracted.

The following fragment demonstrate the models’ attempt to replicate Byron’s characteristic style:

Hence from my native land—lost without trace
Of hills, or wilderness scarce discernible,
Here Adeline descended with her brother;
The deadly wind and damp torrent bore

His motionless bark, unheeded by Heaven:
Through thickets and crevices on the shore
The dim mists seemed little changed by the spray;
The quitting star was little screened by the cloud,
Nor the rising pang by the approaching night!

Similarly, models' attempt at Percy Bysshe Shelley's style:

It is not blasphemy to hope that Heaven
More perfectly will give those nameless joys
Which throb within the pulses of the blood
And sweeten all that bitterness which Earth
Infuses in the heaven-born soul. O thou,
Whose dear love gleamed upon the gloomy path
Which this lone spirit travelled, drear and cold,
Now swiftly lead me to those awful limits
Which mark the bounds of Time and of the Space
Where Time shall be no more.

Phase 2: Controlled Compositions (GPT-3). In **Chapter 4**, the introduction of the summary-poem pair methodology combined with the larger GPT-3 model allowed us to move beyond fragments to complete, structurally integral compositions. Crucially, this approach enabled the decoupling of content from form. To demonstrate this capability, we present a single content summary below—describing a tragic casualty—followed by two distinct realizations of that same narrative: one rendered in the free-verse style of Walt Whitman and another in the rhythmic, rhymed style of Rudyard Kipling.

Input Summary used for generation:

This poem tells the story of a man who was seriously injured after falling from an engine and being dragged along the metals. People who witnessed the accident knew it was hopeless and so they covered him up and left him. His sweetheart visits him on the eve of their marriage and is visibly upset. She departs without saying a word and shortly after, he passes away.

Summary source: *Casualty* by William Ernest Henley.

GPT-3 Davinci fine-tuned: “Casualty”, *replicating the style of Walt Whitman*

UP from the wreckage, up from the bloody mire and the dirt,
Up from the silence of the road, where he lay distinct and broken,
 mistaken for the dead,
Dragging the heavy limbs, the bruised body, back to the threshold
 of the light,
Enduring the fever, the sharp bone, the ambulance moving,
The white room, the antiseptic smell, the nurses moving like angels
 in the dimness.

I see the sweetheart, the betrothed, arriving in the quiet of the evening,
(O the bitter irony! O the ring, the feast, the vows prepared for
 tomorrow!)
She stands by the bed, she is shaking, her face is a mask of
 terror and fear,
She clutches the railing, she kisses the damp hair, she wets his hand
 with her wet tears.
He looks upon her with a glazing eye, a final look, sweet
 and terrible,
The spirit fluttering, the chest heaving its last breath, the silence
 coming in.

Gone! The soul disembodied—the wedding-bed changed for the
 cold bed,
The lover left standing, the hand unclasped, the future snapped
 like a dry leaf,
And she, the mourner, remaining—enveloped in the vast, peculiar
 emptiness of loss,
Breathing the air of the night, the solemn reality, the dull, aching
 silence of the grave.

GPT-3-Curie fine-tuned: “Casualty”, *replicating the style of Rudyard Kipling*

He knew well what shattered the signal-box—
The buffers crashed beneath the wheel,

Having well earned their long repose—
O man of tender heart, to think of thee
Sympathetic tears rise in mine eyes.

The engines, blocked at either end, you know,
Made the air intensely hot for him—
With sparks and cinders thronging down,
And fires that would not be denied.

Fire came from under the car, and fire
Fell on the car from out the night,
Till rivets popped within our sight;
O man of triple steel, to think of thee
Is simply now to weep.

Down on the metals he was cast, and stayed
Broken upon the lines, I fear,
With never a chance to turn or flee—
Who else would have loved thee, Queen
Of all who saw such sights that night?

As if a friend, they covered him—
First with a coat, quite decently—
And then with other makeshift shrouds—
Respectfully awaiting his end
Along the mile of metals.

Now plainly she will speak, now plainly
Smile, now plainly break the bread—
For, losing all, his spirit fled
To Paradise, though he is dead,
And that is what she understands.

Phase 3: High-Fidelity Generation (GPT-4o). By **Chapter 7**, the focus shifted to comparing the quality of AI generation against human benchmarks. The state-of-the-art models (Claude-3-Opus and

GPT-4o) demonstrated high levels of fluency and semantic depth, producing poems that LLM evaluators often rated as comparable to human works.

Claude-3-Opus: “The Mountain Endures”

Silent sentinel, your stony face has seen
The rise and fall of kings and gods and men.
When Joshua’s horns rang out, you stood unmoved,
And watched impassively when Caesar bled.
From your majestic peak, the cosmos wheeled -
Millennia marked by slow, imperceptible change.
What is the span of mortal life to you?
Our greatest deeds and tragedies just wisps
Of cloud that briefly veil your timeworn crags
Before the ceaseless winds disperse them all.
Seven valiant souls now plant their flag
Upon your crown, and stake their claim to fame.
But you remain, as ever, granite-hewn -
A stoic sentry at the gates of time,
Bearing witness to the folly and the dreams
That define the human enterprise.

GPT-4o-mini fine-tuned: “From Golden Walls to Rooted Prayers”

Once I prayed for gold, and higher than my head
Strong walls I built against the battering years;
I did not pray for bread, and went unfed.
Now I pray for strength, and the huge weight of cares
Lies light - as light as dew - on my broad shoulders;
Strength is a golden staff on which the strong lean.

I prayed for glory; the world inscribed my name
In the vital honey of song. In the dark years
My name was nursed on the lips of men, and fed
To be lust and poison and a serpent’s flame.
Now silent I pray; the name all men intone
Kindles no fire, nor fades upon their lips.

Then I prayed for love, of love the only food -
Rimmed joy and crown and instinct of my days.
I had it, for it too is a passionate fire:
Now I pray for a mind as smooth as a root,
Its sap in harmony with the flowering heavens.
I bear the scar of the love I will not slay.

1.6 Structure of the Thesis

This thesis is structured chronologically, following the progression of research conducted between 2021 and 2025, adapting to the rapid advancements in Large Language Model technology. The chapters are organized as follows:

- **Chapter 1: Introduction.** This chapter outlines the research focus on LLM-based poetry generation and evaluation, details the original contributions, describes the thesis structure, lists related publications, and establishes the context of conducting research within the rapidly evolving LLM landscape.
- **Chapter 2: Background.** This chapter provides the necessary theoretical and technical foundations. It covers the evolution of Large Language Models from early concepts to modern architectures like the Transformer, discusses theoretical perspectives on computational creativity and evaluation frameworks, reviews the history of digital poetry generation both before and during the LLM era, examines evaluation methods, and discusses the implications of rapid LLM evolution for research validity.
- **Chapter 3: Training GPT-2 for Romantic-Era Poetry Generation.** This chapter details the initial experiments focused on fine-tuning GPT-2 models to replicate the style of Romantic-era poets. It explores multi-faceted evaluation techniques including BLEU for plagiarism detection and LLM-based classifiers (using BERT) for stylistic assessment.
- **Chapter 4: Fine-Tuning GPT-3 For Poetic Style Generation.** Building on the previous chapter, this chapter introduces a novel methodology using summary-poem pairs to fine-tune GPT-3 models. This approach aims to generate poetry with user-controlled content while maintaining stylistic fidelity. It also further develops LLM-based classifiers (using GPT-3) for automated evaluation.
- **Chapter 5: Assessment Of Zero-Shot And Many-Shot Prompting.** This chapter investigates whether advancements in prompting techniques with early GPT-3.5 and GPT-4 models could replicate specific authorial styles (using Walt Whitman as a case study) without fine-tuning,

ultimately highlighting the limitations of prompting alone for this specific task at that technological stage.

- **Chapter 6: LLM-Based Evaluation Of Poetry.** Shifting focus to evaluation, this chapter introduces a novel methodology inspired by the Consensual Assessment Technique (CAT). It utilizes state-of-the-art LLMs (Claude-3-Opus, GPT-4o) as judges to evaluate human-written poetry, demonstrating their ability to outperform non-expert human judges in aligning with ground truth quality rankings.
- **Chapter 7: Human Poetry vs AI Poetry: LLM Evaluation.** This chapter applies the developed generation and evaluation methodologies to compare the quality of AI-generated poetry (using both prompting and fine-tuning of state-of-the-art models) against human-written poetry benchmarks. It assesses the quality achievable by different AI methods and examines the consistency and potential biases of LLM evaluators.
- **Chapter 8: Conclusions.** This final chapter summarizes the key contributions of the thesis, reflects on the evolution of the methodologies employed, distills durable methodological insights, discusses limitations, and proposes directions for future research in the field of computational poetry generation and evaluation.

Following the main chapters, the Bibliography lists the references cited throughout the thesis.

1.7 Publications Derived from this Research

The research presented in this thesis has contributed to the following publications and presentations, corresponding to different stages of the work:

- **Chapter 3:**
Sawicki, P., Grzes, M., Jordanous, A., Brown, D. and Peeperkorn, M. (2022). Training GPT-2 to represent two Romantic-era authors: challenges, evaluations and pitfalls. In *Proceedings of the 13th International Conference on Computational Creativity (ICCC'22)*.
- **Chapter 4:**
Sawicki, P., Grzes, M., Góes, L.F., Brown, D., Peeperkorn, M., Khatun, A. and Paraskevopoulou, S. (2023). On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)*.

- **Chapter 5:**

Sawicki, P., Grzes, M., Góes, F., Brown, D., Peeperkorn, M. and Khatun, A. (2023). Bits of grass: Does GPT already know how to write like Whitman? In *Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)*.

- **Chapter 6:**

Sawicki, P., Grześ, M., Brown, D. and Góes, F. (2025). Can Large Language Models Outperform Non-Experts in Poetry Evaluation? A Comparative Study Using the Consensual Assessment Technique. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP'2025)*.

Research presented in Chapter 7, is currently under preparation for submission.

1.8 Research Context and Trajectory: Navigating the Rapidly Evolving LLM Landscape

A significant contextual challenge of this research lies in the unprecedented pace at which Large Language Models (LLMs) have evolved during the period of study (2021-2025). This evolution presented both methodological challenges and unique research opportunities that have fundamentally shaped the approach and contributions of this thesis.

When this research began in early 2021, GPT-3 had recently been introduced as a groundbreaking advance in language modeling, while fine-tuned GPT-2 models represented an accessible means for researchers to explore domain-specific text generation. By the time the later chapters of this thesis were being completed, models like Claude-3-Opus and GPT-4o had dramatically expanded the capabilities of zero-shot prompting and sophisticated analysis, challenging earlier assumptions about the necessity of fine-tuning or the limitations of AI in nuanced creative tasks.

This rapid technological progression necessitated continuous adaptation of our research methodology. Our investigation naturally evolved through distinct phases, mirroring the advancements in the LLM landscape:

- **Phase 1: Style Preservation and Content Control with Early to Mid-Generation Models (Chapters 3-5).** Our initial research focused on the core challenge of generating poetry that faithfully replicated authorial style. Chapter 3 explored fine-tuning GPT-2 models, revealing critical challenges like overfitting and plagiarism. Chapter 4 advanced to fine-tuning GPT-3, introducing a novel methodology using ‘summary-poem’ pairs to achieve content control alongside stylistic consistency. Chapter 5 then critically assessed the zero-shot and few-shot prompting capabilities

of early GPT-3.5 and GPT-4 models for stylistic generation, highlighting their limitations for sophisticated style emulation at that stage.

- **Phase 2: Quality Evaluation and Advanced Generation with State-of-the-Art Models (Chapters 6-7).** The advent of highly capable models like Claude-3-Opus and GPT-4o prompted a shift. The emphasis moved towards assessing the overall quality of generated poetry and leveraging these advanced models not just as generators but also as sophisticated evaluators. Chapter 6 introduced a novel LLM-based evaluation framework inspired by the Consensual Assessment Technique (CAT). Finally, Chapter 7 applied these developed methodologies to compare AI-generated poetry against human-written benchmarks, analyzing the quality achievable with the latest models. Crucially, this phase also involved uncovering important evaluator biases, questioning the nature of LLM-perceived quality, and highlighting the distinction between AI's assessment of human versus AI-generated works.

Rather than pursuing a fixed technical approach throughout the research period, we embraced an evolving strategy that:

1. **Systematically tracked and documented** the changing capabilities of successive LLM generations, analyzing how these advances affected creative text generation and evaluation (as demonstrated across Chapters 3-7).
2. **Consciously distinguished** between technically contingent findings (tied to specific model architectures or transient capabilities) and more durable methodological and conceptual contributions. This was crucial for ensuring research validity in a fast-moving field where specific technical implementations can quickly become outdated.
3. **Prioritized the development of conceptual frameworks and evaluation methodologies** designed to maintain relevance despite changing underlying model capabilities. Key durable contributions stemming from this approach include:
 - *Robust Evaluation Frameworks:* The multi-faceted evaluation techniques developed for early models (Chapter 3), including considerations for plagiarism, and the CAT-inspired LLM-as-judge methodology (Chapter 6) offer adaptable frameworks for creative assessment.
 - *Clear Problem Formulation:* Defining stable objectives for computational poetry (such as stylistic fidelity, content control, and quality assessment) provides consistent benchmarks against which future systems can be measured.
 - *Structured Data Paradigms for Generation:* The 'summary-poem' pair approach (Chapter 4) offers a transferable data structuring technique for achieving controllable creative generation,

applicable beyond specific models.

- *Emphasis on Comparative Analysis:* Consistently comparing across models (e.g., GPT-2-small vs. GPT-2-medium, GPT-3-Ada vs. GPT-3-Curie, vs. GPT-3.5-turbo vs. GPT-4, etc.), methods (e.g., fine-tuning vs. prompting), and evaluators (e.g., human vs. LLM, Claude-3-Opus vs. GPT-4o) provides relative insights that remain informative even as absolute capabilities change.

This adaptability allowed our research to remain relevant despite the changing technological landscape. While specific code implementations or model fine-tuning recipes might become outdated as more advanced models emerge, the conceptual approaches to generation control, stylistic analysis, quality evaluation, and the comparative understanding developed herein are intended to provide a durable foundation for future work in computational poetry and creative AI.

Finally, while establishing these frameworks, it is crucial to acknowledge the inherent limitations of using Large Language Models as evaluators of creative work. As discussed in Chapter 7 (Section 7.7), LLMs are not objective arbiters of literary merit; they exhibit distinct evaluation biases and a potential preference for ‘LLM-native’ characteristics—such as high fluency and structural coherence—over the subtler nuances of human artistry. Consequently, the high quality scores assigned by these models to AI-generated content should be interpreted with caution. The automated evaluation methodologies presented in this thesis are therefore proposed not as definitive replacements for human literary expertise, but as scalable, consistent instruments for quality screening that function within their own specific interpretive boundaries.

The chapters of this thesis thus form a chronological progression that not only addresses the research questions but also documents this rapid evolution, providing valuable insights into how researchers in fast-moving technological domains can develop methodologies that produce lasting contributions.

1.9 Conclusion

This chapter has outlined the primary research challenges addressed in this thesis: the development of controllable LLM-based poetry generation and the establishment of robust, scalable evaluation methodologies. We have defined the scope of the investigation, detailed the original contributions, and situated the work within the context of a rapidly evolving technological landscape. Having established these foundations and the research trajectory, the following chapter will provide the necessary theoretical and technical background, reviewing the evolution of Large Language Models and the history of computational creativity that underpins the experiments presented in this thesis.

Chapter 2

Background

This chapter provides comprehensive background information necessary to understand the research presented in this thesis. It begins with an introduction to Large Language Models and word embeddings, detailing their evolution from recurrent neural networks to transformer architectures. The chapter then explores theoretical perspectives on creativity and evaluation frameworks, followed by a historical overview of digital poetry generation before the LLM era. Next, it examines contemporary LLM-based approaches to poetry generation, with a dedicated section on poetry evaluation methods that examines both human and automated assessment approaches.

2.1 General Introduction to Large Language Models

Large Language Models (LLMs) represent a transformative advancement in artificial intelligence, enabling machines to process, understand, and generate human language with unprecedented sophistication. These powerful neural network models are trained on vast corpora of text data, learning to predict and generate sequences of words by capturing complex patterns, semantic relationships, and linguistic structures. LLMs have revolutionized natural language processing (NLP) by demonstrating remarkable capabilities across diverse tasks—from translation and summarization to creative writing and complex reasoning—often approaching or even exceeding human performance in specific contexts. Their effectiveness stems from several key innovations: massive scaling of model parameters (ranging from millions to billions), sophisticated neural architectures optimized for language processing, and increasingly refined training techniques. As we explore the technical foundations of these systems, we begin with one of their fundamental building blocks: word embeddings, which transform discrete linguistic units into continuous vector spaces that capture meaningful relationships between words.

2.1.1 Introduction to Word Embeddings

Word embeddings are a fundamental component of modern natural language processing (NLP) systems. They provide a way to represent words as dense vectors in a high-dimensional space, enabling machine learning models to understand and reason about the relationships between words based on the contexts in which they appear. By capturing semantic and syntactic information, word embeddings form the foundation for a wide range of NLP tasks, from language modeling and machine translation to sentiment analysis and question answering.

This section explores the concept of word embeddings, with a particular focus on the distinction between static and dynamic approaches. We will examine the characteristics, popular models, and applications of each type, comparing their strengths and weaknesses.

2.1.2 Static Word Embeddings

Static word embeddings are a method for representing words as fixed vectors in a multi-dimensional space, based on their overall context within a corpus. In this approach, each word is assigned a single vector representation, regardless of its specific usage in a given sentence. Words that appear in similar contexts across the corpus will have similar vector positions, capturing general relationships between words.

One of the most widely used models for generating static word embeddings is Word2Vec, introduced by (Mikolov 2013; Mikolov et al. 2013). Word2Vec uses a shallow neural network to learn word representations by predicting a target word based on its surrounding context words, or vice versa. The resulting word vectors capture semantic and syntactic similarities, such that words like “king” and “queen” or “man” and “woman” end up close together in the vector space.

Another popular model is GloVe (Global Vectors for Word Representation), developed by Pennington, Socher and Manning (2014). GloVe combines the benefits of global matrix factorization and local context window methods, learning word vectors by optimizing a global objective that considers both word co-occurrence statistics and local context information.

Static word embeddings have proven useful in a variety of NLP applications, such as text classification (Mikolov, Yih and Zweig 2013; Joulin et al. 2016), named entity recognition (Collobert et al. 2011; Lample et al. 2016; Yadav and Bethard 2018), and part-of-speech tagging (Collobert et al. 2011; Huang, Xu and Yu 2015; Plank, Søgaard and Goldberg 2016). By providing a consistent representation of words across different contexts, static embeddings enable models to capture general word relationships and meanings.

However, static word embeddings also have limitations. They lack context sensitivity, meaning that a word like “bank” will have the same vector representation whether it refers to a riverbank or a financial

institution. This can lead to ambiguity and inaccuracies in tasks that require a more nuanced understanding of word meanings based on specific contexts.

2.1.3 Dynamic Word Embeddings

To address the limitations of static word embeddings, researchers developed models that generate dynamic, context-specific vector representations. In this approach, the embedding for a word adapts based on its surrounding words in a sentence, thereby capturing nuanced meanings and resolving ambiguities inherent in polysemous words—words with multiple, distinct meanings (McCann et al. 2017; Peters et al. 2018). Pioneering work like ELMo (Embeddings from Language Models) demonstrated the power of using deep, bi-directional language models to produce such contextualized embeddings (Peters et al. 2018).

Subsequent and highly influential architectures, such as the Transformer (Vaswani et al. 2017), paved the way for models like BERT (Devlin et al., 2018) and the GPT series (Radford et al. 2018, 2019; Brown et al. 2020). Both BERT and GPT models, which are central to the research in this thesis, inherently employ dynamic word embeddings as a core component of their understanding and generation capabilities, as will be detailed in Sections 2.2.5 and 2.2.6.

Dynamic word embeddings have demonstrated significant improvements over static embeddings in various downstream NLP tasks, including sentiment analysis, named entity recognition, and question answering (Peters et al. 2018; Devlin et al. 2018). Unlike static embeddings, which assign a single, fixed vector to each word regardless of its context, dynamic embeddings generate representations conditioned on the surrounding words. This contextual awareness enables NLP models using dynamic embeddings to achieve higher accuracy and a deeper understanding, especially when analyzing complex sentence structures and resolving word-sense ambiguities.

While the context-dependent nature of dynamic embeddings requires on-the-fly computation for each input sentence—contrasting with the pre-computed and stored approach of static embeddings—advancements in hardware, particularly GPU technology, have made the training and deployment of these models increasingly efficient and feasible. As a result, dynamic embedding models have become a cornerstone of modern NLP.

In summary, word embeddings have transformed how NLP models understand and process natural language. While static word embeddings capture general word relationships, dynamic embeddings generate rich, context-specific representations that enable a more accurate and nuanced understanding of language, forming the foundation for the sophisticated LLMs discussed in this thesis.

2.2 Large Language Models

Over the past few years, the field of natural language processing (NLP) has undergone a significant transformation, largely due to the emergence of Large Language Models (LLMs). These powerful models are trained on enormous volumes of text data, enabling them to achieve impressive results across various language-related tasks, including machine translation, text summarization, question answering, and dialogue generation. The journey of LLMs has been marked by gradual enhancements, with each successive architecture building upon the strengths of its predecessors while addressing their limitations. This section will explore the evolution of LLMs, starting from the early stages of recurrent neural networks, progressing through the initial versions of transformer-based models GPT and BERT, and culminating in the most advanced models of today, such as GPT-4, Claude, and Gemini.

2.2.1 Sequential Models: RNNs, LSTMs, and the Challenge of Context

Early neural network approaches for processing sequential data, like natural language, centered around Recurrent Neural Networks (RNNs) (Hopfield 1982; Rumelhart, Hinton and Williams 1986; Elman 1990). Unlike feedforward networks, RNNs possess connections that loop back, allowing them to maintain an internal “memory” or hidden state that captures information from previous steps in the sequence. This design enables them to process inputs like sentences one token at a time, updating their internal state to reflect the sequence processed so far. .

However, simple RNNs faced significant difficulties in learning dependencies between elements far apart in a sequence, a problem often attributed to vanishing or exploding gradients during training (Bengio, Simard and Frasconi 1994). To address this, more sophisticated recurrent architectures were developed, most notably Long Short-Term Memory (LSTM) networks (Hochreiter 1997) and, later, Gated Recurrent Units (GRUs) (Cho et al. 2014). These models incorporate internal gating mechanisms that learn to control the flow of information, allowing them to selectively retain relevant context and forget irrelevant details over much longer sequences than simple RNNs. LSTMs and GRUs became the standard for many NLP tasks for a period.

Despite these improvements in handling longer dependencies, a fundamental characteristic remained: both RNNs and their gated variants process information **sequentially**. This step-by-step computation inherently limits parallelization, making them computationally expensive and slow to train on the increasingly long sequences encountered in large datasets.

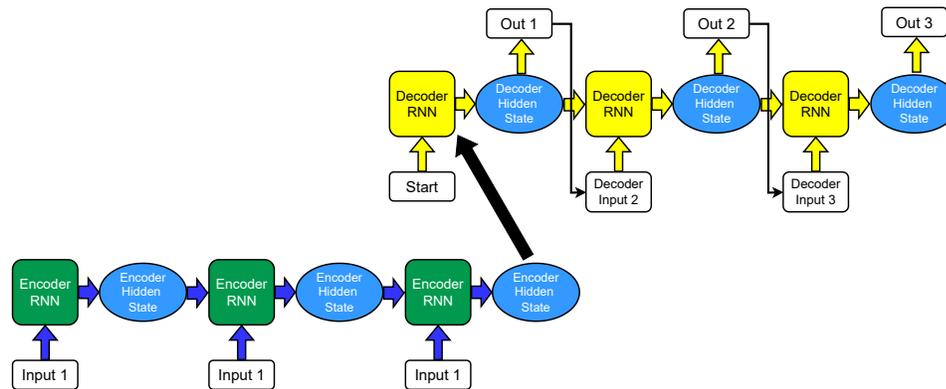


Figure 1: Overview of the Sequence-to-Sequence (Seq2Seq) or Encoder-Decoder model.

2.2.2 The Sequence-to-Sequence Framework and its Bottleneck

A significant advancement for tasks involving mapping an input sequence to an output sequence of potentially different length (e.g., machine translation, summarization) was the Sequence-to-Sequence (Seq2Seq) framework (Sutskever, Vinyals and Le 2014; Cho et al. 2014). Typically implemented using RNNs or LSTMs, Seq2Seq models consist of two main components (see Figure 1):

1. **Encoder:** Reads the entire input sequence, processing it step-by-step and compressing its information into a single, fixed-size **context vector** (the final hidden state of the encoder).
2. **Decoder:** Takes the context vector as its initial state and generates the output sequence one token at a time, conditioning each step on the context vector and the previously generated tokens.

Seq2Seq models achieved state-of-the-art results on many tasks. However, the reliance on a **single fixed-size context vector** proved to be a critical bottleneck. This vector had to encapsulate the meaning and nuances of the entire input sequence, regardless of its length. For long and complex inputs, this compression inevitably led to information loss, hindering the model’s ability to accurately generate corresponding outputs, especially those requiring detailed knowledge of specific parts of the input. This fundamental limitation of the fixed context vector directly motivated the development of attention mechanisms, which aimed to provide the decoder with more flexible access to the entire input sequence during generation.

2.2.3 Attention Mechanisms

Despite LSTM’s superior performance compared to conventional RNNs, it still faces significant challenges when processing long sequences. This limitation stems from the necessity to compress inputs of varying

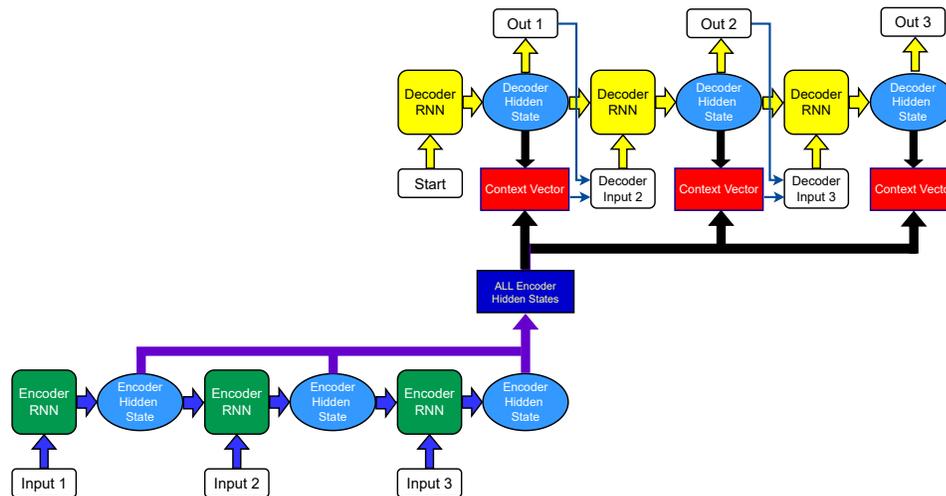


Figure 2: Overview of the Sequence-to-Sequence (Seq2Seq) or Encoder-Decoder model with Attention mechanism.

lengths into fixed-size vectors, a constraint that becomes particularly problematic when handling input and output sequences that exceed the length of training set sequences (Cho 2014).

Traditional Encoder-Decoder architectures operate by generating and updating hidden states sequentially during the encoding process, with only the final hidden state being transmitted to the Decoder. This single-vector bottleneck prompted the development of more sophisticated approaches to sequence processing.

A breakthrough solution to this limitation was introduced by Bahdanau, Cho and Bengio (2014) and subsequently enhanced by Luong, Pham and Manning (2015) through the introduction of the “Attention” mechanism. This innovation, building upon the Encoder-Decoder framework, fundamentally transformed the information flow between encoding and decoding processes.

The Attention mechanism’s key innovation lies in its transmission of all Encoder hidden states to the Decoder, rather than just the final state. During decoding, these vectors, or a subset thereof (Luong, Pham and Manning 2015), undergo a scoring process that enables the Decoder to selectively focus on the most relevant elements, implementing a dynamic attention allocation system, effectively “attending” to the most pertinent information for each decoding step (see Figure 3).

The mechanism operates through a series of well-defined steps (Bahdanau, Cho and Bengio 2014):

1. Initial word vector embedding and LSTM encoding of the input sequence
2. Calculation of alignment scores between encoder outputs and the current decoder state
3. Generation of attention weights through softmax normalization
4. Context vector computation via element-wise multiplication of encoder states and attention weights

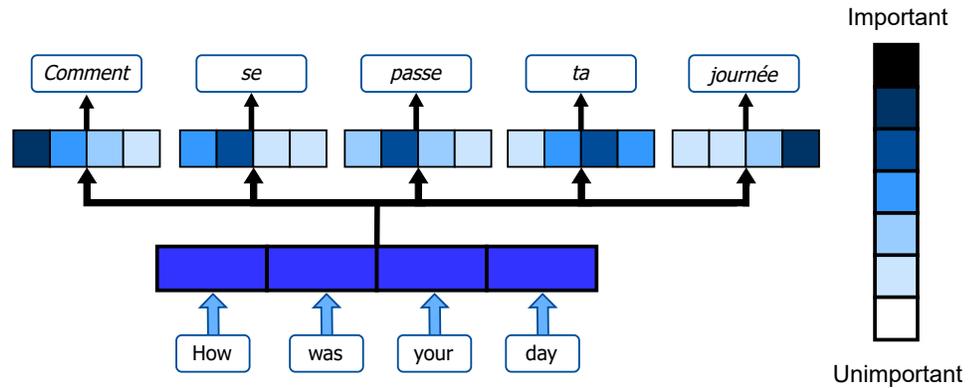


Figure 3: Attention mechanism - scoring

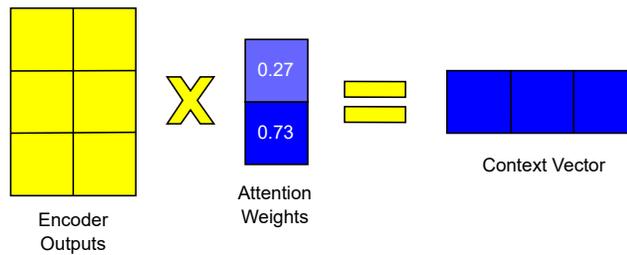


Figure 4: Attention mechanism - context vector.

(see Figure 4).

5. Integration of the context vector with the previous decoder output for new token generation (see Figure 5).

This process iterates until the complete output sequence is generated. Notably, several variants of the Attention mechanism have been developed, as documented by Luong, Pham and Manning (2015), each offering distinct advantages for specific applications.

The Attention mechanism’s versatility is demonstrated through its various implementations, including both cross-attention between input and output elements, as exemplified in the original Transformer model for machine translation (Vaswani et al. 2017), and self-attention within Encoder or Decoder blocks, as utilized in architectures like GPT and BERT, which will be examined in subsequent sections.

2.2.4 The Transformer Architecture

The introduction of the Transformer architecture by Vaswani et al. (2017) marked a revolutionary advancement in sequence-to-sequence modeling by completely eliminating the sequential processing inherent in RNN and LSTM-based approaches. This fundamental shift represents the Transformer’s primary

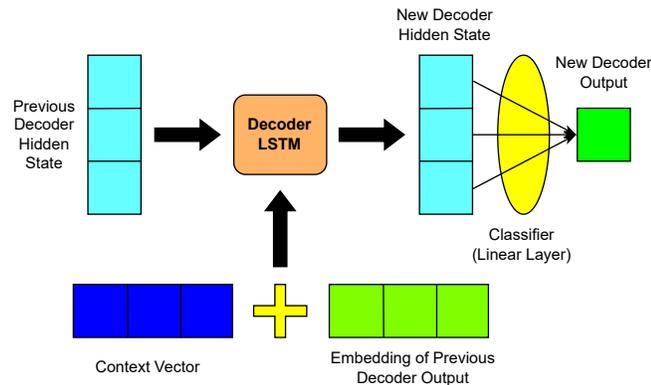


Figure 5: Final step of Attention mechanism.

innovation: rather than processing tokens one after another in sequence, Transformer models process entire sequences simultaneously through self-attention mechanisms, enabling dramatically improved parallelization and more efficient training on modern hardware. This departure from recurrent processing addresses the key limitations of RNNs and LSTMs, particularly their difficulty in capturing long-range dependencies and their resistance to parallelization. By processing all positions at once through attention mechanisms, Transformers can directly model relationships between any words in a sequence, regardless of their distance from each other. Building on this foundation, the Transformer architecture incorporates several additional innovations, including multi-headed attention, where attention operations are performed in parallel across different representation subspaces. Each attention “head” learns to focus on different aspects of the input, capturing various types of relationships and dependencies. The outputs from these multiple attention heads are then concatenated and linearly transformed to produce the final attention output. To maintain information about token positions despite the removal of sequential processing, Transformers employ positional encodings that are added to input embeddings. This combination of parallelized processing, multi-headed attention, and positional encodings has made the Transformer architecture highly efficient, scalable, and the dominant paradigm for natural language processing tasks.

The Transformer’s Encoder comprises multiple identical stacked layers, each containing two primary sub-components: a multi-headed self-attention mechanism followed by a simple fully connected feed-forward layer. Each sub-component is followed by layer normalization (Ba 2016; Vaswani et al. 2017). The multi-headed attention mechanism allows for parallel processing of attention computations, enhancing the model’s ability to capture different aspects of the input sequence simultaneously. Similarly, the Decoder consists of stacked identical layers but incorporates an additional sub-layer that performs attention operations over the Encoder’s output (see Figure 6).

The Transformer implements attention in three distinct ways:

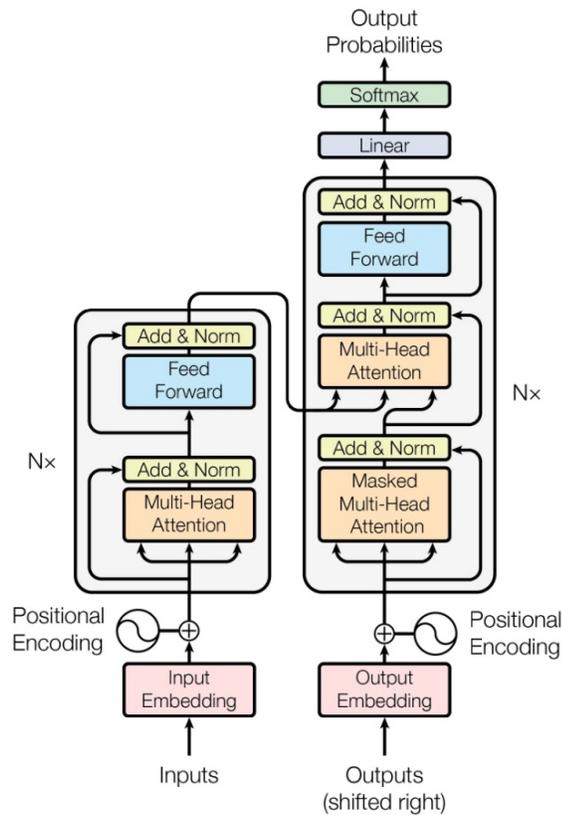


Figure 6: Transformer model architecture (Vaswani et al. 2017)

1. Encoder Self-Attention: Enables each layer to attend to all outputs from the previous layer
2. Decoder Self-Attention: Allows each decoder layer to process all previous layer outputs
3. Encoder-Decoder Attention: Facilitates attention between decoder layers and encoder outputs, similar to the attention mechanism in LSTM-based models (Bahdanau, Cho and Bengio 2014).

To maintain sequence order information without relying on recurrent connections, the model incorporates positional encodings into both Encoder and Decoder inputs. See Figure 6.

The original Transformer implementation features six stacked layers in both Encoder and Decoder components, with embedding vectors of dimension 512. These parameters were empirically determined by Vaswani et al. (2017).

While the original Transformer architecture (Vaswani et al. 2017) introduced both an encoder and a decoder stack for sequence-to-sequence tasks like machine translation, the flexible nature of its core components—particularly the self-attention mechanism and the “transformer layer” (comprising self-attention and feed-forward networks)—has led to its adaptation into several distinct model types. As detailed by contemporary deep learning literature (e.g., (Prince 2023; Bishop and Bishop 2024)), these Transformer-based models are commonly categorized based on the architectural components they employ:

- **Encoder-Decoder Models:** These utilize the full, original two-stack architecture, with an encoder processing the input sequence and a decoder generating the output sequence, connected via cross-attention. They remain prevalent for tasks like machine translation and summarization.
- **Encoder-Only Models:** These models, such as BERT (Devlin et al. 2018), use only the encoder stack of the Transformer. They are typically pre-trained to build rich bidirectional representations of input sequences and are highly effective for natural language understanding tasks like text classification and named entity recognition.
- **Decoder-Only Models:** These models, exemplified by the GPT family (Radford et al. 2018), employ only the decoder stack with masked self-attention to ensure autoregressive behavior. They excel at language modeling and text generation tasks, predicting subsequent tokens based on preceding context.

Understanding these architectural distinctions is crucial, as different model types leverage the Transformer’s capabilities in ways tailored to specific objectives. This thesis primarily focuses on decoder-only models (GPT variants) for poetry generation, but also utilizes encoder-based models (BERT) for certain evaluation tasks in Chapter 3.

The Transformer model (Vaswani et al. 2017) achieved state-of-the-art performance on the WMT 2014 English-to-German and English-to-French translation tasks, surpassing previous models, including ensembles, while requiring significantly less training time. This foundational architecture, in its various forms, underpins both the BERT and GPT models central to our research. As highlighted, these models differ in their specific application of the Transformer’s components: BERT utilizes an encoder-only approach, while GPT employs a decoder-only stack. The following section examines the GPT model family in more detail.

2.2.5 GPT Models

Based on the decoder component of the original Transformer architecture (Vaswani et al. 2017), GPT (Generative Pre-Trained Transformer) models excel at text generation due to their autoregressive nature, making them particularly suitable for creative tasks like the poetry generation explored in this thesis. Unlike the full encoder-decoder structure of the original Transformer designed for sequence-to-sequence tasks, GPT language models employ a decoder-only architecture. This architectural choice, focusing solely on the decoder stack with masked self-attention, is well-suited for language modeling – predicting the next token in a sequence given previous tokens. This approach is exemplified in the original GPT paper (Radford et al. 2018) and is similar to the generative model described in Liu et al. (2018).

GPT models are pre-trained using a “semi-supervised” approach. The model is trained on large corpora of text with the objective of language modeling, which means the system learns to predict the output given the joint probability of the previous sequence representation. Since the text is unlabeled, this resembles an “unsupervised” approach. However, since the result of the prediction is known (i.e., the next token in the text), it can be considered a “supervised” approach. The second, optional stage is fine-tuning the pre-trained model for a specific task using a relatively small amount of labeled or unlabeled data. This approach is similar to using pre-trained word embeddings for various downstream NLP tasks. Instead of ‘learning’ the language from scratch, the models take advantage of the ‘knowledge’ already encoded in the word embeddings.

As explained in Section 2.2.3, the attention mechanism in the Decoder block can be applied either to the hidden states of the Encoder, or it can be applied within the Decoder sequence, in which case it is referred to as “Self-Attention”. Since GPT models have only one stack of Transformer layers, the Attention mechanism they apply is naturally the “Self-Attention”. Additionally, during pre-training and fine-tuning, the Attention mechanism is prevented from “attending” to tokens on the right, which makes it look somewhat similar to sequential processing by RNNs. This is achieved by the “masking” mechanism: after Attention has produced the scores for all the tokens in the input sequence, the scores for the tokens that we want to mask are replaced with a very large negative number, thus ensuring that these tokens will not be selected as outputs. This is done to prevent the model from knowing the tokens that it is supposed to predict. This approach makes the model **auto-regressive**, meaning it predicts the next token in a sequence based on the preceding tokens. During generation, the model uses the complete output of the current step as input to the next step, sequentially producing text one token at a time until a stopping condition is met. Because of that, the GPT models are referred to as “unidirectional”. It has to be noted however, that this “unidirectionality” is different from LSTM-based left-to-right networks in a way that Masked Self-Attention at each step processes the whole available context simultaneously, while LSTMs process the sequence strictly sequentially.

Fine tuning methods for GPT differ according to the task. In order to fine tune the model to generate text in a specific style, for example poetry of Lord Byron, all we need to do is to fine tune it on a dataset of collected works of the author. Moreover, we can fine tune the model on the dataset composed of the works of several authors, where each author’s work will be marked with a respective token. In such a case, during text generation we will have to provide the desired token to force the model to generate text in the style of that specific author.

For other NLP tasks, like text classification, entailment, similarity, or multiple choice question answering, a few additional layers may have to be added.

For example, in multiple choice questions answering, where we are given the context and a set

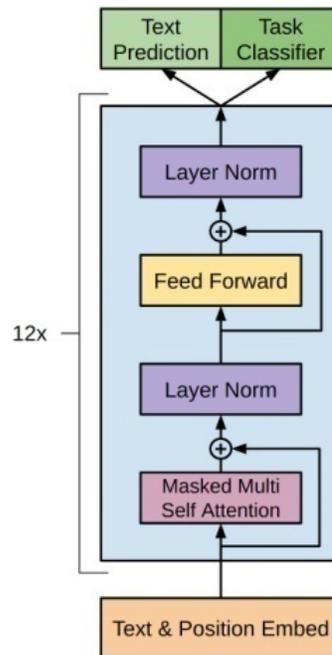


Figure 7: Transformer architecture adapted to various NLP tasks (Radford et al. 2018)

of possible answers, we concatenate the instance of the context with each possible answer, adding $\langle |START| \rangle$ token at the beginning, $\langle |DELIM| \rangle$ token in between the context and the proposed answer, and the $\langle |EXTRACT| \rangle$ token at the end. Each of these samples will be processed independently through the transformer and then a linear layer, and at the end the outputs of the linear layers will be processed through Softmax function to obtain the probability distribution over possible answers. Structures for other tasks are shown in Figure 7 and 8.

Since its inception, the GPT (Generative Pre-trained Transformer) model has undergone significant advancements in terms of scale and training data. The first GPT model (Radford et al. 2018), with 117 million parameters, was trained on a modest 5GB text dataset. In 2019, GPT-2 (Radford et al. 2019) expanded the model's capacity to 1.5 billion parameters and utilized a substantially larger 40GB dataset for training. The introduction of GPT-3 in 2020 (Brown et al. 2020) marked another milestone, with 175 billion parameters and a 570GB training dataset. Despite these substantial increases in size and training data, the core architecture of the GPT model remained largely unchanged, with the primary modifications being the addition of more layers and attention-heads to accommodate the increased scale. Subsequent major advancements focused not only on scale but also on improving model controllability and interaction capabilities, as exemplified by ChatGPT and its successors (Section 2.2.8).

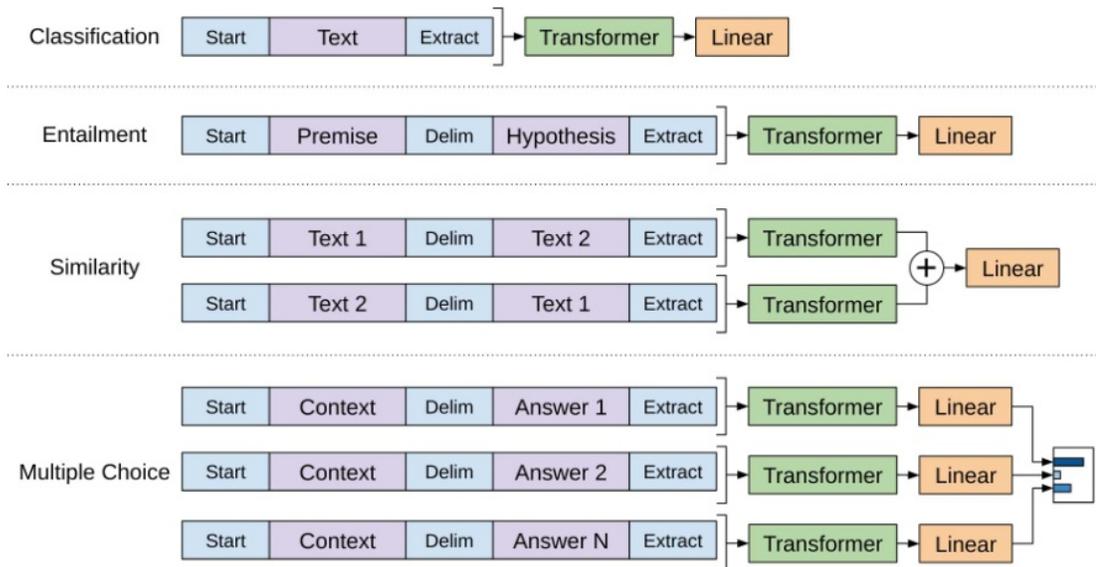


Figure 8: Adapting the pre-trained GPT Transformer architecture for various NLP tasks during fine-tuning. Minimal task-specific layers are added to leverage the model’s learned representations for classification, entailment, similarity, and multiple-choice problems (Radford et al. 2018).

2.2.6 BERT and Its Variants

In contrast to the generative focus of GPT models, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018) leverages the encoder stack of the Transformer architecture. This design makes BERT exceptionally powerful for language understanding and representation learning, rather than direct text generation. Its ability to process entire input sequences *simultaneously* via self-attention, thereby capturing context from both left and right (making it “bidirectional”), underpins its effectiveness in tasks like classification and evaluation (and in that character it was used for stylistic assessment performed in Chapter 3 of this thesis). Unlike GPT’s masked self-attention for autoregressive prediction, BERT’s pre-training employs different objectives, notably the Masked Language Model (MLM) and Next Sentence Prediction (NSP) (Figure 9). In MLM, the model learns to predict randomly masked tokens in the input, a process distinct from, though superficially resembling, the Continuous Bag of Words (CBOW) approach (Mikolov 2013) used with earlier sequential models.

Masked Language Model objective masks randomly 15% of the training sequence and the model has to learn to predict masked tokens. Additionally, since the [MASK] token is unlikely to appear in the training data, if the given token was selected to be masked, in 80% instances it will be replaced with the actual [MASK] token, in 10% instances it will be replaced with a random token from the vocabulary and in another 10% instances it will be left unchanged. The 15% masking ratio was arrived to empirically by

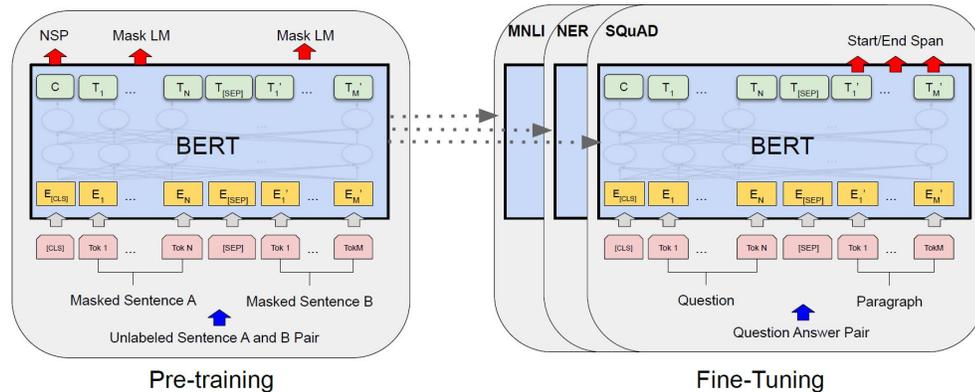


Figure 9: BERT training objectives (Devlin et al. 2018)

the authors (Devlin et al. 2018).

Next Sentence Prediction is a task where the model is given two sentences (A , B) and it is expected to answer whether B is a continuation of A . Since the training data for BERT consists of large amount of unlabeled, continuous texts, creating such training objective is trivial - the dataset simply has to be split into two-sentence fragments. During training, in 50% instances sentence B is actually the continuation of A , and in other 50% instances sentence B is a random sentence from the dataset. This training improves BERTs' ability to understand the relationships between sentences, which aren't necessarily captured by the conventional language modeling objective.

Apart from those differences, BERT is very similar to GPT models. BERT and GPT follow the same semi-supervised approach. The authors of BERT have intentionally released two versions with the numbers of parameters roughly matching those of GPT-2 small and GPT-2 medium, to facilitate the comparison between those two models.

The use of term "bidirectional" is a little confusing here. BERT is not "bidirectional" in the same sense that a LSTM-based network could be, i.e. processing the input from left to right and then from right to left and then concatenating the resulting vectors, as is done in the Embeddings from Language Models (ELMo) (Peters et al. 2018). BERT instead processes the whole input sequence simultaneously.

At the time of its introduction - 2018 - BERT achieved state-of-the-art results on a wide range of NLU tasks. Its success has inspired numerous variants and extensions, such as RoBERTa (Liu et al. 2019), ALBERT (Lan et al. 2019), and XLNet (Yang et al. 2019), each introducing improvements to the original architecture or training procedure.

2.2.7 Adapting Pre-trained Language Models: Fine-Tuning and Prompting

Once Large Language Models like GPT and BERT are pre-trained on vast amounts of text, their extensive learned knowledge can be adapted for specific downstream applications. Two primary paradigms, often used in conjunction or as alternatives, have become central for this purpose, both of which are key to the methodologies explored in this thesis: fine-tuning and what is commonly referred to as “prompting” for in-context learning with base models.

Fine-tuning involves taking a pre-trained model and continuing its training (i.e., updating its weights) on a smaller, curated dataset specific to the target task. For example, a model might be fine-tuned on a corpus of legal documents for legal text analysis, or on summary-poem pairs for controlled poetry generation (as explored in Chapter 4). This process modifies the model’s internal parameters, further specializing its capabilities and often teaching it to respond effectively to specific input formats or instructions relevant to the new domain or task. Fine-tuning can be applied to both generative models like GPT and encoder-based models like BERT, tailoring them for various downstream tasks. While a fine-tuned model still requires an input prompt to initiate its specialized behavior (e.g., a summary to generate a poem, or a sentence for classification), the model is now conditioned to perform optimally on that task with potentially simpler or more structured prompts compared to what a base model might require.

In contrast, **prompting as in-context learning** (often simply called “prompting” when discussing base pre-trained models) aims to guide the behavior of an existing, *unchanged* pre-trained model without altering its parameters. This is achieved by meticulously crafting an input prompt that provides instructions, context, and potentially examples of the desired output (known as zero-shot, few-shot, or many-shot prompting, depending on the number of examples given). The model then generates a completion based on this prompt, leveraging its general pre-trained knowledge and the immediate context provided. This form of prompting has become particularly prominent with very large models that exhibit strong instruction-following and few-shot learning capabilities directly from their pre-training.

It is important to note that even fine-tuned models are “prompted” at inference time; they require an input to generate an output. However, the role and nature of this prompt differ. For a fine-tuned model, the prompt often aligns with the data structure or task format it was specialized on, and the fine-tuning process itself has optimized the model to respond to such prompts. For a base pre-trained model, “prompting” is the primary mechanism for eliciting a desired behavior *without* parameter updates, relying on the model’s ability to generalize from the prompt’s content.

The relative effectiveness, optimal use cases, and interplay between fine-tuning (which shapes how a model responds to subsequent prompts) and sophisticated prompting techniques for base models have evolved significantly with increasing model scale and architectural advancements. The trade-offs and

synergies between these two adaptation strategies are a recurring theme, explored experimentally in the context of poetry generation and evaluation in Chapters 4, 5, and 7 of this thesis.

2.2.8 ChatGPT and onwards

Building upon the remarkable capabilities of GPT-3, OpenAI's journey toward more controllable and human-aligned language models led to several crucial developments. Following GPT-3's success in general language tasks, OpenAI developed InstructGPT (Ouyang et al. 2022) and the GPT-3.5 series (including text-davinci models), which focused on following human instructions more reliably. This evolution culminated in ChatGPT (GPT-3.5-turbo), which represents a significant advancement in making large language models more accessible and useful for everyday interactions, and subsequently led to the even more capable GPT-4 model (Achiam et al. 2023).

ChatGPT's architecture and training process builds directly upon these predecessors, particularly the GPT-3.5 model, while introducing crucial innovations in human feedback and control. While GPT-3 demonstrated impressive capabilities in open-ended text generation, ChatGPT addresses many of its limitations through a sophisticated training process that incorporates Reinforcement Learning from Human Feedback (RLHF) (Knox and Stone 2011). The primary goal of RLHF is to better align the model's outputs with human preferences and intentions, making it more helpful, truthful, and less likely to generate harmful or unwanted content. This approach enhances human control over the model's output and provides a more interpretable ranking of results, moving beyond the purely predictive nature of earlier GPT models.

ChatGPT's training process consists of three distinct stages that build upon GPT-3 and InstructGPT, integrating multiple sophisticated components to overcome limitations of earlier models that primarily relied on next-token prediction. In the first stage, demonstration data is collected to fine-tune the GPT-3.5 model with carefully annotated data (see Figure 10), training a supervised policy that aligns model outputs with human intent. The second stage involves training a reward model (RM) using ranking data generated by human annotators, incorporating Reinforcement Learning from Human Feedback (RLHF) for alignment with human preferences. In the third stage, the policy is optimized using Proximal Policy Optimization (PPO) (Schulman et al. 2017). This combination allows ChatGPT to maintain the powerful language understanding capabilities of GPT-3 while exhibiting more controlled and contextually appropriate responses.

Prompting plays a vital role in both training and using ChatGPT, building upon the few-shot learning capabilities first demonstrated in GPT-3. Prompt learning (Liu et al. 2023) enables the model to directly perform tasks according to human instructions without the need for fine-tuning or changing its parameters based on downstream tasks. This approach represents a significant evolution from the basic prompting techniques used with earlier GPT models, allowing ChatGPT to be more flexible and adaptable to various

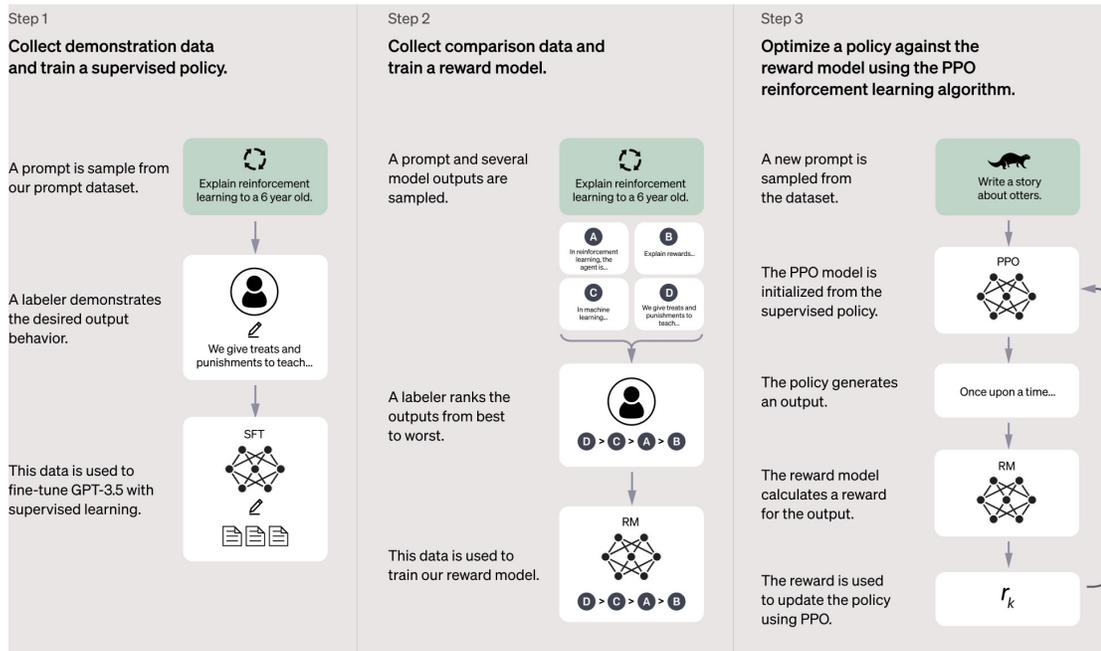


Figure 10: Training process of ChatGPT (OpenAI 2022).

tasks and domains.

The success of ChatGPT paved the way for GPT-4 (Achiam et al. 2023), which represents another significant leap forward in both scale and capability. Released in 2023, GPT-4 builds upon the RLHF techniques refined in ChatGPT while introducing multimodal capabilities, allowing it to process both text and images. GPT-4 demonstrates substantially improved performance across a wide range of tasks, including more complex reasoning, creative writing, and technical analysis, while maintaining the controllability and alignment achievements of ChatGPT. Furthermore, GPT-4 and subsequent models like GPT-4o featured significantly larger context windows compared to their predecessors, enabling them to process and reason over much longer prompts and documents. This increase in context capacity is directly relevant to the feasibility of the many-shot prompting experiments (Chapter 5) and the in-context evaluation methods (Chapters 6 and 7) explored later in this thesis.

Alongside GPT-4, other powerful proprietary large language models have emerged, such as Anthropic’s Claude (Anthropic 2024) and Google’s Gemini (Pichai and Hassabis 2025). These models perform on par, if not better than GPT-4 on several benchmarks (Center for Research on Foundation Models 2025).

2.2.9 Open-source LLMs

While the core experiments in this thesis primarily utilize proprietary models from OpenAI and Anthropic due to their state-of-the-art capabilities and accessibility during the research period, the parallel development of powerful open-source LLMs represents an important context, offering alternatives with greater transparency and fostering community-driven innovation. The landscape of Large Language Models (LLMs) has evolved significantly since the initial releases of models like BERT and early versions of GPT. While these pioneering models were open-source, enabling researchers to explore and build upon their architectures, the field has since diverged into two distinct paths: proprietary models developed by large technology companies and a burgeoning ecosystem of open-source alternatives.

A pivotal moment in the open-source movement was the work of Eleuther AI, an organization dedicated to advancing open-source AI research. Eleuther AI released a series of powerful models, including GPT-Neo (Black et al. 2021), GPT-J (Wang and Komatsuzaki 2021), and GPT-NeoX-20B (Black et al. 2022), which provided the research community with access to high-quality LLMs comparable to proprietary counterparts. These models not only demonstrated that state-of-the-art performance could be achieved without proprietary constraints but also fostered innovation by allowing researchers to study, modify, and build upon their architectures freely.

Another significant contribution to the open-source LLM landscape was Meta AI's Open Pre-trained Transformers (OPT) models (Zhang et al. 2022). The OPT series ranges from 125 million to 175 billion parameters, providing the research community with access to models comparable in scale to proprietary systems like GPT-3. By open-sourcing both the model weights and the training code, along with detailed documentation of the training process and evaluation metrics, Meta AI aimed to promote transparency and facilitate reproducibility in large-scale language model research.

The momentum continued with the release of Meta's LLaMA series (Touvron et al. 2023b,a; Dubey et al. 2024), which further showcased that high-quality language models could be developed and shared openly. Building upon these foundations, adaptations and improvements such as Alpaca (Taori et al. 2023) and Vicuna (The Vicuna Team 2023) leveraged instruction-tuning techniques to enhance model performance. Similarly, the BLOOM model (Scao et al. 2023), a collaborative effort led by Hugging Face, marked another significant milestone by offering a multilingual model trained on 46 natural languages and 13 programming languages.

More recent developments include Falcon (Almazrouei et al. 2023), which achieved competitive performance with significantly fewer parameters than its predecessors. Mistral (Jiang et al. 2023) and Mixtral (Jiang et al. 2024) introduced innovative architectural improvements for enhanced efficiency. These models have demonstrated that open-source alternatives can approach, and in some cases match, the performance of proprietary models in specific tasks while requiring substantially fewer computational

resources.

The open-source approach offers several distinct advantages over proprietary models. First, it enables complete transparency in model architecture and training procedures, facilitating reproducibility and scientific scrutiny. Second, it allows for community-driven improvements and adaptations, leading to rapid innovation and specialized applications. Third, researchers and organizations can modify and fine-tune these models for specific use cases without the restrictions typically associated with proprietary solutions.

An important distinction to note is that while open-source LLMs provide transparency in their architecture and code, many still maintain limitations regarding their training data. The Llama 3 series, as detailed in the paper (Dubey et al. 2024), was pre-trained on approximately 15T (trillion) multilingual tokens (compared to Llama 2's 1.8T tokens), with extensive data curation processes including de-duplication, quality filtering, and safety measures. Regarding alignment techniques, Llama 2 notably incorporated RLHF (Reinforcement Learning from Human Feedback) in its development process, while Llama 3 shifts to an approach that employs supervised fine-tuning (SFT) followed by Direct Preference Optimization (DPO) rather than traditional RLHF methods. The paper specifically notes this was chosen for stability and scalability advantages. Earlier models like Alpaca and Vicuna utilized instruction tuning based on synthetic data derived from proprietary models rather than direct RLHF. More recent models such as Mistral and Falcon have adopted various preference-based alignment techniques, though the specific methodologies and data sources are not always comprehensively documented. This partial transparency remains an ongoing challenge in the open-source LLM ecosystem, where code availability does not always translate to full training data or alignment methodology disclosure, a pattern consistent across various open-source models from Llama to Mixtral to Gemma.

However, open-source LLMs face notable challenges. The computational resources required for training and fine-tuning, while lower than those needed for models like GPT-4, remain substantial for many research institutions. Quality assurance and safety considerations also present significant hurdles, as the open nature of development can make it more difficult to implement consistent standards and safeguards (Anwar et al. 2024).

2.3 Theoretical Perspectives and Evaluation Frameworks for Computational Creativity

Poetry generation is widely treated as a creative task, making the study of creativity relevant to our research on LLM-based poetry systems. While computational systems can generate poetry without explicitly modeling creativity, understanding theoretical perspectives on creativity provides valuable context for evaluating and discussing their outputs. This section examines key theoretical frameworks for

conceptualizing creativity and reviews prominent approaches to its evaluation, offering conceptual tools that inform our research on LLM-based poetry generation and assessment. These frameworks help situate our technical work within broader discussions of computational creativity.

2.3.1 The Question of Intentionality in Creative Agency

Before exploring specific frameworks for evaluating creativity, we must address a fundamental question: can computational systems—particularly LLMs—possess genuine creative agency, or are they merely executing processes designed by humans? This question goes beyond whether machines can generate novel and valuable outputs to ask whether they can have intrinsic motivations for creating.

Guckelsberger, Salge and Colton (2017) frame this as the “why?” question in computational creativity. They argue that current systems cannot explain their creative choices without ultimately referring to their designers’ goals. Ada Lovelace articulated this limitation nearly two centuries ago: “The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform” (Menabrea and Lovelace 1843). Despite enormous technological advances, this critique remains relevant. While modern LLMs can generate impressive creative works, their goals and values are externally imposed rather than intrinsically motivated.

To explore this, some researchers draw upon the enactive approach to cognitive science, a perspective emphasizing that cognition arises through dynamic interactions between an embodied agent and its environment, grounded in the agent’s autonomy and self-maintenance (Di Paolo, Buhrmann and Barandiaran 2017; Basu 2004). Within this context, Guckelsberger, Salge and Colton (2017) propose that intentional creative agency, viewed through an enactive lens, requires two key conditions linked to an agent’s precarious existence:

1. **Constitutive autonomy:** The system must generate and maintain its own identity, physically individuating itself from its environment.
2. **Adaptivity:** The system must regulate its interaction with the environment relative to conditions that threaten its viability.

From these conditions emerge two forms of creativity that such systems would necessarily exhibit: autopoietic creativity (modifying structure to ensure continuous existence) and adaptive creativity (responding to perturbations with novel, viability-preserving actions). In this enactive view, creativity is grounded in an agent’s maintenance of its precarious existence—a radical departure from anthropocentric models that treat creativity as a specialized human capability.

This perspective suggests that LLMs fundamentally lack intentional creative agency despite their impressive outputs. While this theory grounds creativity in responding to precarious existence, LLMs

demonstrate no intentional effort to respond or adapt to any such precarity. They do not actively maintain themselves against conditions that might threaten their operational continuity, nor do they face intrinsic existential threats requiring creative self-preservation. Their value functions reflect external human design goals rather than inherent concerns arising from a precarious existence.

However, this limitation does not diminish the practical utility of LLMs as creative tools. Rather, it helps us situate them within a more nuanced understanding of creativity, distinguishing intentional agency from what might be termed “generative creativity”—a system’s ability to create “new patterns or behaviours regardless of the benefit to that system” Bown (2012), as cited in Guckelsberger, Salge and Colton (2017). We might view LLMs through this lens, as exhibiting this form of pattern generation rather than intentional creative agency motivated by intrinsic goals.

The distinction between generating creative outputs and being a creative agent is crucial for understanding and evaluating LLM-based poetry generation. It suggests that we should assess such systems not by whether they possess human-like creative intentions, but by how effectively they serve as tools extending human creative capabilities.

This framework also helps us navigate debates about machine creativity by separating two distinct questions:

1. Can machines produce creative outputs? (Clearly yes, according to the definitions of novelty and value discussed in Section 2.3.2, although precise evaluation remains complex)
2. Can machines be creative agents with their own intrinsic motivations? (More problematic under the enactive framework)

As we proceed to examine specific evaluation frameworks, we should keep in mind this fundamental distinction between creative production and creative agency. The frameworks we discuss—spanning overarching models like Rhodes’ Four P’s (Person, Process, Product, Press) (Rhodes 1961) to more specific approaches such as Jordanous’s Standardized Procedure for Evaluating Creative Systems (SPECS) (Jordanous 2012) and Amabile’s Consensual Assessment Technique (CAT) (Amabile 1983)—primarily address the evaluation of creative outputs and processes rather than the more challenging question of intentional agency.

Familiarity with these frameworks is nonetheless valuable for developing effective LLM-based poetry generation systems. While current LLMs may lack intentional creative agency in the enactive sense, they still require sophisticated evaluation methods to assess the quality, novelty, and value of their outputs. The theoretical frameworks discussed in the following sections provide useful tools for this purpose, even as we acknowledge the philosophical limitations of computational creativity in its current form.

This pragmatic framing sidesteps many of the theoretical challenges raised by Guckelsberger, Salge

and Colton (2017). Rather than asking whether LLMs possess intentional creative agency, we focus on how they can effectively augment human creativity. The question becomes not “Can LLMs be creative in their own right?” but “How can LLMs best serve as creative tools for human poets and researchers?” Evaluating such systems requires different criteria than those used for autonomous creative agents. As Veale (2013) notes, the value of AI creative systems often lies in their ability to provoke new perspectives or provide unexpected inspirations for human creators. Consequently, since the AI lacks intrinsic intentionality, we must view it not as a partner in a symmetrical collaboration, but as a sophisticated instrument. This distinction frames our subsequent exploration of creativity: we are not assessing the ‘mind’ of the machine, but rather the novelty and value of the artifacts it generates under human guidance.

2.3.2 Defining Creativity

The concept of creativity has been approached from various disciplinary perspectives, with definitions evolving significantly over time. Boden (2004) offers a widely accepted definition of creativity as “the ability to come up with ideas or artifacts that are new, surprising and valuable.” This tripartite definition emphasizes that creativity involves not only novelty but also elements of surprise and value.

The question of novelty itself is complex. Boden distinguishes between psychological or “P-creativity” (ideas that are new to the individual who produces them) and historical or “H-creativity” (ideas that are novel within the entire history of human thought). This distinction is particularly relevant for computational systems, which may produce outputs that are novel from the system’s perspective but replicate existing human creations—a challenge sometimes referred to as “reinventing the wheel” (Colton 2008).

Value is equally multifaceted, encompassing aesthetic, emotional, practical, and cultural dimensions. Wiggins (2006) gives a formal framework, which requires an evaluation function for creative systems. It highlights the challenge of operationalizing value computationally, particularly given that value is often inherently contextual and culturally determined. This raises profound questions about whether value can be meaningfully assessed by algorithms or must ultimately rely on human judgment.

Surprise—the third element in Boden’s definition—relates to the violation of expectations and has been linked to cognitive processes involving schema disruption and reformation. For a computational system, implementing surprise requires some model of expectations that can be meaningfully contravened—a significant challenge that goes beyond mere randomness.

Rhodes (1961) proposed an influential framework for understanding creativity through what has become known as the “Four P’s” model:

1. **Person:** The individual characteristics, traits, and cognitive processes of those who create.

2. **Process:** The mechanisms, strategies, and methods employed in creative acts.
3. **Product:** The outputs or artifacts that result from creative endeavors.
4. **Press:** The environmental, social, and cultural contexts that influence creativity and its assessment.

This multifaceted framework recognizes that creativity is not merely a property of individuals or outputs but emerges from a complex interplay of factors (Jordanous 2016). In computational creativity research, these dimensions have guided the development of systems that aim to embody creative capabilities, with varying emphasis on each aspect.

MacKinnon (1970) expanded on this framework, arguing that the “person” dimension encompasses not just cognitive abilities but also personality traits, motivations, and values—aspects that are particularly challenging to model in computational systems. Similarly, the “press” dimension highlights the social embeddedness of creativity, suggesting that true creative systems might need to engage with cultural contexts and receive feedback from communities of practice (Csikszentmihalyi 1999).

Boden (2004) further distinguishes between three types of creativity:

1. **Combinatorial creativity:** Creating new ideas by combining existing concepts in unfamiliar ways.
2. **Exploratory creativity:** Exploring a structured conceptual space of possibilities.
3. **Transformational creativity:** Transforming the conceptual space itself, changing the rules to enable previously impossible ideas.

This taxonomy provides a useful framework for analyzing computational approaches to creativity. Early systems predominantly engaged in combinatorial creativity, assembling existing elements according to predefined rules. More sophisticated approaches, including many neural network-based systems, operate through exploratory creativity, navigating vast conceptual spaces of possibilities defined by their training data and architecture. Transformational creativity—the most profound form in Boden’s hierarchy—remains elusive in computational systems, as it requires the ability to fundamentally redefine the conceptual space itself.

In the context of computational poetry generation, systems primarily engage in combinatorial and exploratory creativity (Gervás 2019). However, as Kantosalo and Toivonen (2016) suggest in their analysis of human-AI co-creativity (distinguishing between “alternating” and “task-divided” approaches), collaboration may offer a pathway towards achieving more transformational outcomes. When human input guides the AI’s adaptation of its conceptual frameworks, computational systems might overcome the limitations that typically constrain them to combinatorial and exploratory modes.

2.3.3 Creativity Evaluation Frameworks

Evaluating creativity presents significant challenges due to its subjective and context-dependent nature. Several frameworks have been developed to address this complexity, each with particular strengths and limitations.

The SPECS Approach

Jordanous (2012) developed the Standardized Procedure for Evaluating Creative Systems (SPECS), which emphasizes contextual, multi-faceted evaluation. SPECS involves three steps:

1. Identifying and articulating what creativity means in a specific domain
2. Deriving standards or criteria based on this definition
3. Evaluating the system against these standards using appropriate metrics and tests

This approach recognizes that creativity manifests differently across domains and contexts, requiring tailored evaluation methods. Through empirical analysis of how creativity is discussed across domains, Jordanous (2012) identified 14 key components that frequently emerge in discussions of creativity:

1. Active involvement and persistence
2. Dealing with uncertainty
3. Domain competence (e.g., understanding poetic forms, meter, rhyme)
4. General intellectual ability
5. Generation of results
6. Independence and freedom
7. Intention and emotional involvement
8. Originality (e.g., novel metaphors, unique thematic treatments)
9. Progression and development
10. Social interaction and communication
11. Spontaneity and subconscious processing
12. Thinking and evaluation
13. Value (e.g., aesthetic appeal, emotional resonance, conveying meaning)

14. Variety, divergence, and experimentation

Jordanous (2012) emphasizes that the relative importance of the 14 SPECS components varies significantly by creative domain. For poetry generation, evaluation literature consistently underscores the centrality of several key aspects that map directly onto these components. Originality (encompassing novel metaphors, unique thematic treatments, or surprising juxtapositions), the assessment of Value (covering aesthetic appeal, emotional resonance, and the successful conveyance of meaning), and the demonstration of Domain competence (including the skillful handling of poetic forms, meter, rhyme, and linguistic quality, as discussed by Manurung, Ritchie and Thompson (2012)) are frequently identified as particularly salient criteria. The artistic and technical demands of poetry make these components critical for judging creative success in this specific field (Jordanous 2012; Lamb, Brown and Clarke 2018).

Demonstrating its practical utility, the SPECS approach has been applied to various computational creativity systems, including poetry generators. Jordanous (2012) herself showed how the framework facilitates systematic comparison of different systems based on such defined criteria, revealing strengths and weaknesses that might not be apparent through more informal evaluation methods.

Creative Tripod

Colton (2008) proposed the “creative tripod” framework, suggesting that for a system to be perceived as creative, it must exhibit three key properties:

1. **Skill:** Technical proficiency in the domain
2. **Appreciation:** The ability to evaluate its own and others’ outputs
3. **Imagination:** The ability to produce surprising and unexpected results

This framework provides a more accessible alternative to complex evaluation models, focusing on core properties that audiences intuitively associate with creativity. It has frequently been employed within the computational creativity community both as an analytical tool to assess the capabilities of existing creative systems across these three dimensions, and as a conceptual guideline for designing new systems that aspire to more holistic creative behaviour. For example, in poetry generation, skill might manifest as grammatical correctness and adherence to formal constraints; appreciation as the ability to select high-quality outputs; and imagination as the generation of novel metaphors or unexpected juxtapositions.

The creative tripod is particularly valuable for its emphasis on appreciation—the capacity for self-evaluation that distinguishes sophisticated creative systems from mere generators. As Colton (2008) argues, a system that can critically assess its own outputs demonstrates a higher level of creative capability than one that produces without evaluation.

Descriptive Models of Computational Creativity

Among the influential descriptive frameworks in the field, the complementary *FACE* and *IDEA* models proposed by Colton, Charnley and Pease (2011) offer a comprehensive language for analyzing creative systems. The *FACE* model describes a creative *act* by detailing the generative processes involved in producing concepts, expressions, aesthetics, and framing information. In parallel, the *IDEA* model assesses the potential *impact* of these acts by situating them in relation to existing knowledge and potential audience reception, categorizing outputs on a spectrum from simple re-invention to genuine discovery. While these frameworks offer a powerful lens for describing the internal processes and potential significance of creative software, the empirical evaluation central to this thesis aligns more directly with techniques that assess the creative product through expert consensus.

Consensual Assessment Technique (CAT)

A gold standard for empirical creativity assessment is the Consensual Assessment Technique (CAT), developed by Amabile (1983). The core principle of CAT is that creativity is best understood not through a fixed checklist of features, but as a subjective quality that is most reliably judged by experts in a given domain. It bypasses the intractable problem of defining creativity by operationalizing it as the consensus judgment of knowledgeable individuals.

The CAT methodology follows a specific procedure to ensure reliable and valid assessments. This typically involves:

- **Expert Judges:** Appropriate experts with extensive domain experience are selected to serve as raters.
- **Independent Evaluation:** Judges evaluate the creative products independently, without consultation, to avoid groupthink.
- **Holistic and Relative Judgment:** Judges are asked to use their own internal, holistic understanding of creativity to rate the products relative to one another within the presented set, rather than scoring them against a predefined, external rubric.
- **Inter-Rater Reliability:** The level of consensus among judges is measured statistically, typically using an Intraclass Correlation Coefficient (ICC). A high level of reliability indicates that the experts agree on which products are more or less creative, validating the assessment.

Due to its empirical rigor, the CAT has been widely adopted in creativity research and has been applied to the domain of computational poetry evaluation, most notably by Lamb, Brown and Clarke (2016). However, its primary limitation is practical: the technique is resource-intensive, requiring significant time

and effort to recruit and coordinate multiple human experts. This makes it challenging to apply at scale, particularly for the large volumes of output that can be generated by modern AI systems.

The primary limitation of CAT—its reliance on costly and time-consuming human expert panels—presents a significant barrier to the scalable evaluation of AI-generated artifacts. This thesis addresses this challenge directly. Chapter 6 proposes and validates a novel adaptation of the CAT methodology, where state-of-the-art Large Language Models (LLMs) are employed as surrogate expert judges. By leveraging the principles of relative, in-context, and holistic assessment, we investigate whether an LLM-based framework can replicate the consensus-driven nature of CAT while offering unprecedented scale and efficiency.

2.3.4 Creativity in Computational Poetry Generation

Poetry generation presents unique challenges for computational creativity systems due to poetry’s multifaceted nature. Manurung, Ritchie and Thompson (2012) propose that poetic competence requires a system to produce outputs that satisfy three key properties:

1. **Meaningfulness:** The poem should convey a coherent message.
2. **Grammaticality:** The poem should adhere to linguistic conventions.
3. **Poeticness:** The poem should exhibit poetic features like meter, rhyme, or metaphor.

Balancing these requirements is particularly challenging, as improvements in one dimension often come at the expense of others (Manurung, Ritchie and Thompson 2012). Indeed, as Gervás (2011) notes, computational poetry systems often excel at formal aspects (poeticness) while struggling with semantic coherence (meaningfulness). This difficulty highlights the complex interplay of form and content that makes poetry particularly challenging for computational systems. While many systems can generate text that satisfies formal constraints, achieving the semantic depth and emotional resonance characteristic of human poetry remains elusive (cf. Lamb, Brown and Clarke 2017; Manurung, Ritchie and Thompson 2012).

Evaluating computational poetry generation involves considering multiple facets of creativity. Drawing from broader psychological and philosophical theories, evaluation approaches often examine dimensions such as the work’s perceived **Novelty** (uniqueness and surprise), **Value** or **Meaning** (appropriateness, usefulness, or the intentional conveyance of ideas), **Craft** or technical skill (effective use of poetic techniques), and the **Reaction** it evokes (emotional response and interest generated). As surveyed by Lamb, Brown and Clarke (2018), these kinds of dimensions align with broader creativity assessment frameworks while also needing adaptation to address the specific nature of poetry as an art form.

An important consideration in computational poetry evaluation is the role of interpretation. Human readers actively construct meaning from poetic texts, often finding significance beyond an author's explicit intentions (Lamb, Brown and Clarke 2018). This interpretive dimension complicates evaluation, as the perceived meaning of a poem may emerge from the reader's engagement rather than being inherent in the text itself. For computational systems, this raises questions about whether the absence of explicit meaning in generated text necessarily indicates a creative failure or might instead create space for reader interpretation.

2.3.5 The Role of Value and Intention: Mere Generation?

A crucial debate in computational creativity concerns the role of intention and value judgment, particularly regarding what distinguishes sophisticated creative systems from those performing only "mere generation." Ventura (2016) addresses this by proposing a spectrum of generative systems, outlining representative levels with increasing complexity and potential for intentionality:

1. **Randomization:** Producing outputs stochastically without regard for constraints or knowledge.
2. **Plagiarization:** Randomly selecting and outputting items from a predefined inspiring set of known artifacts.
3. **Memorization:** Re-representing an inspiring set, potentially introducing unintentional variations or errors during reproduction.
4. **Generalization:** Building a model from an inspiring set and generating novel variations based on learned patterns, potentially guided by bias or regularization.
5. **Filtration:** Generating candidates (e.g., via generalization) and then explicitly selecting the best ones using a predefined fitness or objective function.
6. **Inception:** Enhancing the generation or filtration process by incorporating and leveraging an external knowledge base (e.g., semantic relations, world facts).
7. **Creation:** Grounding the evaluation process in perceptual abilities, allowing the system to assess artifacts based on how they might be perceived or understood in the world.

Ventura's analysis suggests that moving towards the higher end of this spectrum involves increasing sophistication, characterized by the ability to:

1. Generate novel artifacts (often with increasing intentionality regarding novelty).
2. Evaluate the quality or value of generated artifacts (explicitly or implicitly).

3. Incorporate evaluation or knowledge to guide or filter the generation process.
4. Potentially possess mechanisms allowing for explanation or justification of creative choices (highlighted when contrasting generative vs. filtering approaches).

Applying this framework, one might consider that many early poetry generation systems would fall towards the lower end of this spectrum (closer to “mere generation”), while more advanced approaches incorporate elements of intentionality and self-evaluation, particularly when augmented with feedback mechanisms or critical components, placing them higher on the spectrum. This graduated approach allows for more nuanced assessment of computational systems, recognizing incremental progress toward more sophisticated creative capability. Most contemporary poetry generation systems likely operate at the **Generalization, Filtration, or Inception** levels of this spectrum, with true perceptually grounded **Creation** remaining an aspirational goal.

The question of value judgment remains particularly complex. As Colton and Wiggins (2012) argue, creativity requires not just the generation of novel artifacts but the ability to distinguish valuable novelty from mere difference. This raises profound questions about whether computational systems can truly assess value or merely simulate value judgments based on patterns observed in human creative works.

Wiggins (2006) suggests that value judgments involve domain-specific knowledge, making them particularly challenging to implement in computational systems—a difficulty especially pronounced in the pre-LLM era. For poetry generation, for example, effective value judgment might involve not just understanding linguistic patterns but also cultural references, emotional resonance, and aesthetic traditions. While contemporary Large Language Models, trained on vast textual corpora encompassing cultural and aesthetic discussions, demonstrate a significantly enhanced capacity to learn and replicate patterns associated with human value judgments (a capability explored empirically in Chapters 6 and 7 of this thesis), the fundamental challenge persists in nuanced forms. The degree to which these models achieve genuine contextual understanding versus sophisticated pattern matching remains a subject of ongoing research and debate, and capturing subtle, novel, or deeply embodied aesthetic values likely continues to pose difficulties even for the most advanced AI systems.

The role of intention similarly raises complex philosophical questions. As discussed in Section 2.3.1, distinguishing between systems that merely exhibit intentional behavior and those with genuine intentional states is a central challenge for computational creativity (Dennett 1987). This distinction is critical: while poetry generation systems may produce outputs that appear intentional, the question of whether they possess genuine, self-motivated creative intentions remains contentious.

Cohen (2002), discussing his *AARON* program, emphasizes the distinction between generating original output and genuine creativity. He argues that true creative agency requires autonomy, specifically

the ability for the system to develop and modify its own underlying criteria for generating work, not merely modifying rules based on fixed criteria. This perspective aligns with the view that computational creativity should be evaluated on its own terms rather than as a simulation of human creativity.

2.3.6 Evaluation in Practice: Challenges and Approaches

Translating theoretical frameworks into practical evaluation methodologies presents numerous challenges. Common pitfalls identified in the field include:

1. **Evaluation as an afterthought:** Many systems are developed without clear evaluation criteria in mind
2. **Over-reliance on human judgment:** While human assessment is valuable, it can be subjective and inconsistent
3. **Emphasis on products over process:** Many evaluations focus exclusively on outputs rather than creative processes
4. **Lack of comparative frameworks:** Evaluations often fail to situate systems within broader contexts

To address these challenges, researchers have developed various practical approaches to evaluation. Ritchie (2007) proposes an empirical framework based on analyzing the properties of generated artifacts in relation to a “reference” set of human-created works. This approach uses criteria such as typicality (how similar generated works are to the reference set) and quality (how well generated works satisfy domain-specific criteria) to assess creative achievement.

Pérez y Pérez and Sharples (2004) suggest a different approach, evaluating systems based on cognitive plausibility—how well they align with theories of human creative cognition. This perspective values systems that model human creative processes, even if their outputs do not always match human quality standards.

In practical AI poetry evaluation, researchers have employed various methodologies, each with inherent strengths and weaknesses. These include:

1. **Human evaluation studies**, where judges assess generated poems, often comparing them to human-written works (Lamb, Brown and Clarke 2016). While this approach can yield rich qualitative data and nuanced aesthetic judgments, it is often resource-intensive, time-consuming, and can suffer from inter-rater inconsistency and subjectivity.

2. **Metric-based evaluation**, which utilizes quantitative measures of linguistic features or adherence to formal constraints such as rhyme and meter (Oliveira 2017; Agirrezabal, Oliveira and Ormazabal 2023). Such metrics offer objectivity and scalability, making them useful for large-scale comparisons or iterative development. However, they frequently struggle to capture subtle qualitative aspects of poetic achievement, such as emotional resonance or semantic depth.
3. **Turing test approaches**, designed to test whether human judges can distinguish between human-written and computer-generated poems (Köbis and Mossink 2021; Gunser et al. 2022; Rahmeh 2023; Porter and Machery 2024). These provide a clear benchmark for a system’s ability to mimic human output to the point of indistinguishability. Nevertheless, Turing tests face criticism for potentially reducing the assessment of creativity to mere human mimicry, neglecting the underlying generative process or the system’s capacity for appreciation and intentionality, aspects deemed crucial for genuine computational creativity by researchers like Colton and Pease (Colton, Pease and Saunders 2018).

The distinct limitations of each individual methodology highlight the challenge in finding a single, perfect evaluation method for computational poetry. This underscores the value of employing multi-faceted evaluation strategies that combine different approaches to gain a more comprehensive understanding of a system’s creative capabilities.

2.3.7 Implications for Computational Poetry Generation

The theoretical perspectives and evaluation frameworks discussed in this chapter have significant implications for computational poetry generation and evaluation:

1. **Multiple Evaluation Approaches**: The multidimensional nature of creativity suggests that no single metric can capture the quality of generated poetry. Effective evaluation therefore requires multiple complementary methods, incorporating both formal analysis and human assessment.
2. **Style vs. Creativity**: The distinction between style replication and genuine creativity informs approaches to developing poetry generation systems. As Brown and Jordanous (2022) note, generating poems in an existing style involves elements of creativity but raises questions about the boundaries between imitation and innovation. This tension is particularly relevant for LLM-based approaches, which are becoming increasingly capable of stylistic mimicry.
3. **AI-Augmented Creativity**: The limitations of purely computational approaches to creativity highlight the potential value of systems designed to augment human capability. Such interactive

approaches may offer a path to more sophisticated creative outcomes, combining computational capabilities with human aesthetic judgment and cultural knowledge.

4. **Meta-Creativity:** The ability to evaluate poetry relates to meta-creative capabilities—the capacity to assess and reason about creative works. This connects to Colton (2008)’s notion of appreciation as a component of computational creativity. Systems that can not only generate but also critically evaluate poetry represent a more advanced form of creative capability.
5. **Contextual Creativity:** The “Press” dimension of the Four P’s model (Rhodes 1961) highlights the importance of cultural and social contexts in creativity, a view strongly supported by systemic theories emphasizing that creative value is recognized and judged within a specific social field (Csikszentmihalyi 1999). Consequently, poetry generation systems that actively engage with their context—for instance, by incorporating contemporary themes (Rashel and Manurung 2014), leveraging culturally specific knowledge bases (Veale 2013), or adapting to specific audience needs in collaborative settings (Kantosalo et al. 2014)—may achieve a higher level of perceived creative relevance than those operating in isolation.
6. **Process Transparency:** Several evaluation frameworks emphasize the importance of understanding creative processes, not just products (e.g., (Jordanous 2012; Colton 2008)). This suggests value in systems that can explain or document their creative decisions, providing insight into the generative process alongside the final output. While achieving true introspection into the internal mechanisms of large language models remains a significant challenge (e.g., Bommasani et al. (2021)), techniques such as chain-of-thought prompting aim to elicit step-by-step reasoning traces that offer a degree of process transparency (e.g., Wei et al. (2022)). The pursuit of explainability is considered valuable for understanding and trusting creative AI outputs (e.g., Adadi and Berrada (2018)).

2.3.8 Conclusion: Theoretical Foundations for Computational Poetry

The frameworks and approaches discussed in this chapter provide essential context for understanding the evolution of computational poetry generation, which we will explore in the following sections. As we trace the development from rule-based systems to neural approaches and eventually to Large Language Models, these theoretical foundations help us assess not just technical advancements but also progress toward more meaningful creative capability.

The exploration of creativity reveals its complex nature, often conceptualized through elements like novelty, value, and surprise (Boden 2004), and understood via multi-dimensional models such as the Four P’s (Rhodes 1961). While philosophical questions surrounding machine intentionality and agency persist (Guckelsberger, Salge and Colton 2017), a range of frameworks exist to analyze creative

systems. These include descriptive models that characterize creative acts, as well as more procedurally-oriented approaches for empirical assessment, such as SPECS (Jordanous 2012). However, the empirical evaluation central to this thesis draws its primary inspiration from the Consensual Assessment Technique (CAT) (Amabile 1983). CAT, with its reliance on expert consensus to evaluate subjective creative outputs, provides a robust and empirically-grounded foundation that directly informs the novel LLM-based evaluation method developed in this work.

For computational poetry generation specifically, these theories highlight the inherent challenges in balancing formal constraints with semantic depth and achieving outputs perceived as both novel and valuable (Manurung, Ritchie and Thompson 2012). Evaluating success requires nuanced approaches that consider these multiple facets.

These theoretical underpinnings are important not only for understanding the historical evolution of digital poetry (Sections 2.4 and 2.5) but also, as detailed in the following section 2.3.9, for interpreting the specific methodologies developed and results obtained throughout this thesis. They provide a lens through which to analyze the capabilities and limitations of LLMs as tools within the creative process.

2.3.9 Linking Creativity Theory to Thesis Methodology

The theoretical perspectives and evaluation frameworks discussed in this section provide context for understanding and interpreting the practical methodologies and findings presented throughout this thesis. While the work primarily adopts an assistive tool stance (Section 2.3.1), acknowledging the AI primarily as a tool rather than an autonomous creative agent, these theories illuminate several key aspects:

- **Generation Methods and Creativity Types:** The generation methods employed throughout this thesis map onto established creativity frameworks in distinct ways. Both fine-tuning on authorial corpora (Chapters 3 and 4) and prompt-based generation (Chapters 5 and 7) primarily exemplify Boden's *exploratory creativity*. In both cases, the LLMs navigate the vast latent space of language possibilities learned during pre-training. The fine-tuning process (Chapters 3 and 4) narrows this exploration, guiding the model towards patterns characteristic of the specific training corpus (e.g., a particular poet's style), while prompt-based generation (Chapters 5 and 7) uses the input prompt to constrain and direct the exploration within the broader capabilities of the base model.

The summary-poem fine-tuning method introduced in Chapter 4 adds a significant element of *combinatorial creativity*. This technique explicitly requires the model to synthesize two distinct conceptual inputs: the stylistic patterns implicitly learned from the poet's work during fine-tuning, and the novel semantic content and structure dictated by the user-provided summary. This directed combination of existing style elements with new thematic material aligns well with the definition

of combinatorial creativity.

Viewed through the spectrum of generative systems (Ventura 2016), the progression of methods also shows increasing sophistication. Basic fine-tuning (Chapter 3) and the less successful prompting attempts for style (Chapter 5) align well with “Generalization”, where the model learns statistical patterns and produces novel variations. It also likely incorporates “Inception” by demanding the model leverage its broad knowledge base (beyond the fine-tuning data) to meaningfully fulfill the semantic requirements of diverse summaries. Furthermore, the explicit quality-based selection process used in Chapter 7 to identify the best AI poems for comparison is a clear instance of “Filtration”, applying an external evaluation metric (LLM judgment) to select preferred outputs from a generated set. While the underlying creative *type* (exploration, combination) might not fundamentally change between earlier fine-tuning and later state-of-the-art prompting (Chapter 7), the latter demonstrates a vastly more effective form of “Generalization” and “Inception” for achieving high perceived quality, even without explicit fine-tuning for that goal.

- **Promoting Novelty and Assessing Value:** This thesis addresses the core creative requirements of novelty and value through two distinct methodological thrusts. The first is a generative approach aimed at fostering *novelty*. By using summary-poem pairs for fine-tuning (Chapter 4), we move beyond mere style replication towards combinatorial creativity, where the model is guided to produce novel poetic expressions based on user-defined content. This structured approach aims to balance learned patterns with new information, steering the system away from simple reproduction. The second thrust is evaluative, focused on assessing *value*. In Chapters 6 and 7, we develop and apply a framework where LLMs themselves become the arbiters of poetic quality, providing a scalable method for measuring the perceived value of generated works.
- **Evaluation Criteria and SPECS:** The evaluation criteria employed in Chapters 6 and 7 (Creativity, Quality, Innovativeness, Similarity, Poeticness) align with several components identified in Jordanous (2012)’s SPECS framework. For instance, “Quality” relates to SPECS components like *Value* and *Domain Competence*; “Innovativeness” maps to *Originality* and *Variety/Divergence*; “Similarity” relates inversely to *Originality*; and “Poeticness” aligns with *Domain Competence*. The multi-faceted evaluation approach adopted from Chapter 3 onwards also resonates with SPECS’ emphasis on assessing multiple dimensions of creativity.

2.4 The Evolution of Digital Poetry Generation in the Pre-LLM Era

The fusion of technology and creativity has long been a subject of fascination, leading to the development of computational systems capable of generating poetry. Before the advent of Large Language Models

(LLMs), researchers explored various methods to enable machines to compose verses that emulate human poetic expression. This section delves into the landscape of digital poetry generation prior to the LLM era, tracing the evolution from early rule-based systems to more sophisticated approaches that bridged the gap between artificial intelligence and artistic creativity. We will examine the key methodologies, challenges, significant achievements, and prevailing themes that shaped the field during this formative period.

2.4.1 Early Approaches: Rule-Based Systems and Traditional NLP Methods

Rule-Based Poetry Generation

The initial forays into digital poetry generation were predominantly rule-based, relying on handcrafted templates and formal grammatical rules to produce poetic text. These systems were designed to mimic the structural elements of poetry, such as rhyme schemes, meter, and stanza formation.

One notable example is *ASPERA* (Automatic Spanish Poetry Expert and Rewriting Application) proposed by Gervás (2001), which introduced a semi-automatic system combining rule-based reasoning with case-based reasoning to compose formal Spanish poetry. *ASPERA* utilized a knowledge base of literary rules to select appropriate metric structures and adapted examples from a corpus of poems to generate verses. While it maintained metrical correctness, the system faced limitations in semantic coherence.

Similarly, Gervás (2000) presented *WASP* (Wishful Automatic Spanish Poet), a rule-based system focused on generating Spanish poetry by controlling pattern selection and vocabulary usage. The system demonstrated that acceptable verse could be generated with minimal linguistic knowledge, emphasizing the importance of pattern matching and formal poetic rules.

Template-Based and Constraint Satisfaction Methods

Template-based approaches provided a straightforward mechanism for poetry generation by filling pre-defined templates with appropriate words or phrases. These methods often incorporated constraint satisfaction techniques to adhere to poetic structures.

Rashel and Manurung (2014) developed *Pemuisi*, which generated topical poems in Indonesian language based on current news articles using constraint satisfaction to enforce poetic constraints. The system demonstrated success in producing contextually relevant poems, highlighting the potential of combining content sources with formal constraints.

Another system proposed by Toivanen et al. (2013) leveraged answer set programming to generate poetry in English that satisfied various constraints such as rhyme, meter, and syllable patterns. This declarative approach separated the specification of poetic constraints from the exploration of possible

poems, offering flexibility in generation.

Early Natural Language Processing Techniques

Traditional NLP methods were employed to analyze and generate text based on linguistic rules and statistical models. Systems like the one proposed by Tobing and Manurung (2015) utilized dependency parsing and semantic preservation to generate metrically constrained English language poetry from input texts. While the system maintained semantic relationships and metrical form, it faced computational challenges, which were addressed through optimizations like text summarization and word sense disambiguation.

2.4.2 Frameworks for Understanding Computational Poetry Generation

Lamb, Brown and Clarke (2017) developed a comprehensive taxonomy of computer-generated poetry techniques, providing a unified framework for understanding various approaches across artistic, hobbyist, and scientific communities. Their analysis categorized generative poetry systems into three main categories based on the level of computational or human intervention involved: mere generation, human enhancement, and computer enhancement.

Mere generation techniques, such as using templates, Markov chains, or context-free grammars, often produce limited or nonsensical results due to their reliance on basic algorithmic processes. In contrast, enhancement methods offer more sophisticated outcomes by incorporating either human input or advanced computational techniques.

Human enhancement involves direct interaction or editing of computer-generated content, allowing for a more guided and curated creative process. Computer enhancement, on the other hand, utilizes AI techniques like optimization and knowledge representation to improve the quality and coherence of the generated output.

While acknowledging that basic generation techniques may have reached their limits, Lamb, Brown and Clarke (2017) highlight the significant potential for growth in the field through the application of advanced computational techniques. They particularly emphasize the promise of combining semantic understanding with artistic optimization to generate poetry that is both meaningful and well-crafted.

2.4.3 Evolution Towards Sophisticated Approaches

Evolutionary Algorithms and Optimization

Evolutionary algorithms introduced a stochastic element to poetry generation, allowing systems to explore a vast search space of possible poetic constructions. Manurung (2004) presented *MCGONAGALL*, an English language poetry system that framed poetry generation as a state space search problem. By

employing genetic algorithms combined with Lexicalized Tree Adjoining Grammar (LTAG), the system optimized poems for meaningfulness, grammaticality, and poeticness. Although successful in balancing semantic and formal constraints, it highlighted the inherent complexity of multi-objective optimization in poetry generation.

Similarly, to generate English language poetry, Rahman and Manurung (2011) applied the Strength Pareto Evolutionary Algorithm (SPEA2) to better handle the simultaneous optimization of semantics and metrical form, demonstrating improvements over traditional linear combination methods.

Statistical and Machine Learning Methods

The use of statistical models marked a significant shift, enabling systems to learn patterns from large corpora and generate poetry that more closely resembled human language. Barbieri et al. (2012) introduced a method that combined Markov models with constraint satisfaction to generate song lyrics in English that imitated an author's style while meeting structural requirements.

This approach effectively maintained local stylistic properties and global structural constraints. Greene, Bodrumlu and Knight (2010) developed an unsupervised system for analyzing stress patterns in poetry. Their method employed finite-state transducers (FSTs) and utilized EM training – a statistical algorithm (Expectation-Maximization) used in this context to learn the probabilities of mapping words to syllable stress patterns directly from examples of poetry lines and their target meters. This system facilitated both poetry generation in English and poetry translation from Italian into English, preserving metrical patterns across languages.

The challenge of adapting generative techniques to diverse linguistic structures was also a key focus in the pre-LLM era. For instance, Hämäläinen and Alnajjar (2019) tackled the complexities of Finnish morphosyntax, employing genetic algorithms to maintain grammatical correctness and poetic structure in their generative system. Similarly, Oliveira (2012) introduced *PoeTryMe*, a modular and customizable platform that utilized semantic networks and grammar templates initially to generate poetry in Portuguese. Its adaptability across languages and integration of various generation strategies made it a cornerstone in computational poetry, with Oliveira et al. (2017) later adapting *PoeTryMe* for Spanish and English, demonstrating the feasibility of a multilingual approach while also highlighting differences due to resource availability. These examples illustrate efforts to build more versatile and linguistically nuanced poetry generation systems prior to the dominance of large pre-trained models.

Knowledge-Based Approaches

Veale (2013) showcased the effectiveness of incorporating knowledge bases and conceptual metaphors

to generate poetry with depth and intentionality. Veale’s system, *Stereotrope*, relies on a database of stereotypes derived from web similes, normative relationships between stereotypes mined from web questions, and Google n-grams. By leveraging these knowledge sources, *Stereotrope* selects metaphorical conceits, creates conceptual blends, maps properties to emotions, and renders these elements through various poetic tropes. The system also incorporates the “Keats heuristic”, which suggests that aesthetically pleasing forms are perceived as more truthful. *Stereotrope* demonstrates how computational poetry can move beyond simple form manipulation to generate content worthy of poetic expression, while acknowledging the essential role of human interpretation in the poetic process.

Neural Networks and Deep Learning

The advent of neural networks brought a new level of sophistication to poetry generation, enabling the incorporation of complex linguistic features. A notable example is the work by Zugarini, Melacci and Maggini (2019), who developed an innovative syllable-based neural language model using LSTMs to generate Italian poetry mimicking Dante Alighieri. Critically, their approach captured metrical structure and style without relying on poetic training data. They processed poetry at the syllable level—recognizing syllables as fundamental regulators of poetic form, meter, and rhyme—and employed a multi-stage transfer learning process: first learning from modern Italian, then refining with Dante’s non-poetic works, and finally training on the *Divine Comedy*. During generation, outputs were automatically scored on formal elements (tercet structure, hendecasyllabic meter, rhyme, vocabulary) to select the most authentic examples. While this syllable-based architecture effectively replicated Dante’s style, often deceiving non-experts, expert judges noted limitations in emotional depth. Nevertheless, this work demonstrated that focusing on fundamental poetic units like syllables could allow neural models to capture complex stylistic elements implicitly, challenging traditional rule-based approaches.

Zhang and Lapata (2014) developed a model that jointly performed content selection and surface realization for Chinese classical poetry. By capturing representations within and across poem lines, the RNN-based system effectively generated quatrains that respected structural and aesthetic requirements.

Lau et al. (2018) introduced a neural system called *Deep-speare* that jointly modeled language, meter, and rhyme to generate quatrains of Shakespearean sonnets. These models demonstrated proficiency in formal elements but revealed challenges in achieving readability and emotional impact. Examples of generated quatrains are shown below:

pours out its distant margin’s towering steep
and winding river where a torrent rides
as on his swift way on a rushing tide

high in the leafy murmur of the strain

when as we made him that triumphant lies
 whose beauty like a front of pomp, no doubt
 can wear upon her head? what is the world
 which lays it to be great, and now is mine

a lover in my spirit's sad distress
 here i rejoice, and call us at a hour
 which must regret its follies first our last
 and told the future, to the one's to give

in darkness to behold him, with a light
 and him was filled with terror on my breast
 and saw its brazen ruler of the night
 but, lo! it was a monarch of the rest

These examples illustrate *Deep-speare*'s strong command of formal poetic elements (meter and rhyme) while demonstrating the semantic and emotional limitations identified by the researchers.

Emphasis on Coherence and Meaning

Researchers increasingly focused on enhancing semantic coherence and meaningful content, recognizing that poeticness is not solely determined by form. Wang et al. (2016) proposed a Planning based Poetry Generation (PPG) system that combines content planning with generation to create coherent and semantically consistent Chinese classical poems. The system first plans the sub-topics for each line using either an RNNLM (Recurrent Neural Network Language Model)-based or knowledge-based method, then generates the poem line by line using a modified attention-based RNN encoder-decoder framework. Human evaluation results showed that the PPG system outperformed baseline models in terms of poeticness, fluency, coherence, and meaning, with nearly half of the machine-generated poems being indistinguishable from human-written ones.

InkWell: Exploring Consciousness and Creativity in Digital Poetry

InkWell, developed by Gabriel (2016), advanced digital poetry by generating haiku-like poems using a hybrid of symbolic AI and machine learning. Presented at a writers' conference, its poems were initially

believed by the participants to be human-written, demonstrating AI’s potential for human-like creativity.

Below are examples of *InkWell*’s poetry:

deep in the dark—
the power of snow
walking in the deepness

the powerful head
designates its powerful head
to support cognition

scrupulous in the twilight—
the price of gold chases
the way of the world in power

The study emphasized balancing poetic constraints such as sense, tone, form and emphasis. Conference participants sought evidence of human consciousness not only in the text but also through interaction with the author, suggesting that the Turing test involves both imitation and probing for consciousness. One particularly striking example that garnered praise was:

time of life issue:
a bird of prey pulls up
out of the way into the palm

InkWell’s success in producing nuanced poetry relied on reasoning about word meanings and context. This work highlights AI’s capability to generate poetry exhibiting qualities of human consciousness and underscores the importance of human perception in evaluating creative AI systems. Collectively, these sophisticated approaches in the pre-LLM era demonstrate a persistent engagement with core generative challenges: achieving a delicate balance between adherence to poetic form (meter, rhyme) and the creation of meaningful, coherent content; imbuing generated text with semantic depth and creativity beyond mere pattern replication; and navigating the complexities of different linguistic structures, as exemplified by efforts in Finnish and multilingual systems. While significant progress was made, these inherent difficulties in generation also underscored the critical need for robust methods to assess the outputs, a topic explored next.

2.4.4 Evaluating Computational Poetry

Evaluating computational poetry poses significant challenges due to the subjective nature of aesthetic appreciation. Research by Lamb, Brown and Clarke (2015) highlighted these difficulties when investigating non-expert judges using structured evaluation metrics on human-written poetry. Their experiments revealed a counter-intuitive finding: non-experts consistently preferred amateur poetry over professional works, rating the former higher on criteria such as quality, skill, and typicality. This “reversed pattern” was hypothesized to stem from the inaccessibility of complex professional poetry to untrained readers. While a follow-up experiment using more accessible children’s poetry removed this specific reversal, it still failed to reliably differentiate between amateur and professional creations. Lamb, Brown and Clarke (2015) thus demonstrated that structured metrics do not obviate the critical need for domain expertise in judges, cautioning that evaluations relying on non-experts or those with unclear qualifications may yield unreliable or even misleading results.

Building on this, Lamb, Brown and Clarke (2016) explored the use of the Consensual Assessment Technique (CAT) (Amabile 1983) with quasi-expert judges (graduate students in Experimental Digital Media) for evaluating digital poetry. Although overall inter-rater reliability was lower than typically expected for CAT, the judges achieved good consensus on the most and least creative poems in the set, while disagreeing more on mid-range works. Through qualitative analysis of the judges’ justifications, the study identified four recurring desiderata implicitly used in their assessments: Reaction (the emotional and interest response evoked), Meaning (the coherence and intentional conveyance of ideas), Novelty (the uniqueness and surprising aspects), and Craft (the skillful and effective use of poetic techniques). This work suggested limitations of applying standard CAT to heterogeneous digital poetry but also proposed these desiderata as potential domain-specific criteria, emphasizing the value of combining such product-based insights with process-based evaluation methods.

Collectively, these studies underscore that evaluating poetry, whether human or machine-generated, is complex and highly dependent on the evaluator’s expertise. They demonstrate limitations both when non-experts use structured metrics and when quasi-experts apply consensus techniques to diverse digital works, pointing towards the need for nuanced, multi-faceted evaluation approaches sensitive to both the creative product and the process behind it.

2.4.5 Interactive Systems and Creative Support

Interactive poetry generation systems have increasingly focused on augmenting the user’s creative process with AI assistance. Kantosalo et al. (2014) developed the *Poetry Machine*, a co-creative system designed to help children write poetry by overcoming the “blank page” problem. Using a fridge magnet metaphor

for manipulation, the system provided draft poems consisting of poetry fragments that users could modify. Their evaluation framework, detailed in Kantosalo, Toivanen and Toivonen (2015), demonstrated how different interaction patterns affected the creative process, revealing that computational aids can help overcome the initial barriers of creative writing for young users.

The development of interactive systems that support human creativity became increasingly prominent. Oliveira, Mendes and Boavida (2017) and Uthus et al. (2019) emphasized user control and collaboration, aligning AI tools with the creative practices of poets.

Zhipeng et al. (2019) exemplified the move towards systems that augment human creativity with *Jiuge*. By allowing users to interact and refine generated Chinese classical poems, *Jiuge* demonstrated the potential of AI as an assistant rather than a replacement in the creative process.

2.4.6 Conclusion: Evolution of Digital Poetry Generation

The pre-LLM era of digital poetry generation laid a crucial foundation, evolving from rule-based attempts at formal mimicry to more sophisticated neural approaches striving for semantic depth. Throughout this period, researchers persistently grappled with fundamental generative challenges: achieving a delicate balance between poetic form and meaningful content, ensuring semantic coherence and genuine creativity beyond mere pattern replication, and adapting techniques to diverse linguistic contexts. Systems like *PoeTryMe* exemplified the drive for modularity and adaptability, while early neural networks began to hint at the potential for capturing more nuanced aspects of poetic expression. Concurrently, the intricate problem of evaluating poetic quality spurred methodological innovation. This era of exploration, marked by both significant achievements in human-AI collaboration and a clear understanding of the inherent complexities, ultimately set the stage for the transformative impact of Large Language Models, which promised to leverage vast linguistic knowledge to address these enduring challenges in new ways.

2.5 Digital Poetry Generation in the LLM Era

2.5.1 Introduction

The application of Large Language Models (LLMs) to poetry generation marks a significant shift from pre-LLM era techniques. While earlier systems grappled with balancing rule-based structures and semantic coherence (as discussed in Section 2.4), LLMs bring vast pre-trained linguistic knowledge, offering new potentials and distinct challenges for creating poetic text. This section reviews examples of how LLMs have been employed for poetry generation, examining foundational models, approaches to specific poetic constraints, advanced techniques, and the evolving role of human-AI collaboration in this new paradigm.

LLMs, such as GPT-2, GPT-3, GPT-4, and BERT, have demonstrated remarkable capabilities in generating coherent and contextually appropriate text. Their application to poetry generation leverages these strengths to produce verses that mimic human poetic styles. Researchers have employed these models to generate various forms of poetry, including sonnets, limericks, haikus, rap lyrics, and classical Chinese poetry. The primary goal is to create systems that can (semi)autonomously produce poetic content that is indistinguishable from human-authored poetry, both in form and substance, and as current research shows, the field is approaching that goal (Porter and Machery 2024).

2.5.2 Foundational Models and Approaches

GPT-Based Models

The GPT series has been the cornerstone of many poetry generation systems. GPT-2 and GPT-3, with their ability to generate fluent and contextually relevant text, have been fine-tuned for poetry generation tasks. For example, fine-tuning GPT-2 on Emily Dickinson’s poetry allowed the model to mimic her characteristic punctuation and capitalization, although it struggled to capture the deeper philosophical themes (Dai 2021). Similarly, Chapter 3 of this thesis explores fine-tuning of GPT-2 models to generate poetry in the styles of specific authors and Chapter 4 focuses on fine-tuning GPT-3 models to approximate not only the style of specific authors but also the user controlled content.

Early explorations into fine-tuning GPT-class models for controlled creative text generation also include work on domains similar to poetry, such as song lyrics. Wesek (2021), in a detailed study, investigated fine-tuning GPT-2 for conditional lyrics generation. By employing special tokens to encode metadata like genre, author, and year, and through systematic experimentation with dataset sizes, input construction strategies (including token type and position IDs), and output sampling techniques, Wesek (2021) demonstrated methods for achieving fine-grained stylistic control over generated lyrics. This work highlighted the practical considerations and effectiveness of targeted fine-tuning for guiding GPT-2’s creative output based on structured conditional inputs, an approach conceptually aligned with efforts to control poetic style and content.

BERT and Masked Language Models

BERT’s strong bidirectional language understanding, particularly through its Masked Language Model (MLM) pre-training, makes it effective for tasks involving text manipulation and transformation, rather than solely end-to-end generation. The MLM feature allows BERT to predict masked words based on surrounding context, a capability useful for revising existing texts.

Oliveira (2021) demonstrated this by using a Portuguese BERT model to creatively transform song

lyrics. Content words in the original lyrics were masked, and BERT provided replacement candidates. Instead of using the most probable suggestion, candidates were selected based on their semantic similarity (using BERT embeddings) to a specified theme, enabling controlled shifts in meaning. Oliveira (2021) also experimented with adding morphological and metrical constraints to the selection process. While showcasing BERT’s potential for guided creative transformation, this work highlighted the inherent trade-offs between achieving strong theme alignment or adhering to constraints and maintaining grammatical fluency. BERT’s language understanding capabilities have also been utilized for evaluating poetry, as detailed in Section 2.6.1.

2.5.3 Addressing Specific Poetic Constraints and Forms

Controlling stylistic and structural elements in poetry generation—such as rhyme schemes, meter, sentiment, and specific poetic forms—remains a significant challenge. To address this, Shao et al. (2021) proposed masking self-attention mechanisms to associate multiple attributes during generation, demonstrating their approach on Chinese classical poetry with simultaneous control of style and sentiment. Similarly, Li et al. (2020) introduced a conditional generation method with symbolic control for rigid formats, incorporating special symbols to guide structural constraints across both Chinese SongCi and English Sonnets. Both approaches successfully maintain format accuracy while preserving poetic qualities, with particularly strong results for Chinese poetry generation.

Despite these advancements, imitating the style of specific poets remains complex. Chapter 5 of this thesis demonstrates that base GPT models, including the latest GPT-4, when used without fine-tuning, failed to replicate Walt Whitman’s style accurately, even with extensive prompting, highlighting the limitations of current models in capturing nuanced authorial voices without specific training.

Further highlighting the difficulties in achieving precise formal control through simple prompting, Agirrezabal and Oliveira (2024) conducted a quantitative analysis of the open-source Llama2 model’s ability to adhere to metrical constraints. When prompted in a zero-shot fashion to generate poems with specific numbers of lines, stanzas, and syllables per line, their findings, measured using the automated evaluation system *Erato* (Agirrezabal, Oliveira and Ormazabal 2023), revealed significant struggles. Llama2 was often unable to consistently meet these formal demands, particularly concerning syllable counts and less common metrical patterns. This research underscores that while LLMs can generate fluent and poetic-sounding text, precise adherence to formal structural constraints often remains elusive without more targeted interventions beyond basic prompting, a challenge ironically inverse to traditional rule-based systems which could handle meter but struggled with semantics.

Addressing constraints also involves specialized techniques tailored to particular languages or poetic forms. For languages with complex character systems like Chinese, token-free models operating at

the character level have proven effective. For instance, Yu et al. (2024) proposed *CharPoet*, which generates classical Chinese poetry character by character, ensuring strict adherence to poetic formats while maintaining content relevance. This character-level focus, sometimes combined with iterative polishing frameworks as shown by Deng et al. (2020), helps meet the rigid constraints typical of classical Chinese poetry.

Specialized forms like rap lyrics present unique challenges due to their emphasis on rhythm, dense rhyme schemes, and cultural context. Researchers like Nikolov et al. (2020) and Wentink (2023) have developed systems specifically aimed at generating coherent rap verses that address these demands. Beyond specific forms, progress has also been made in multilingual contexts. For example, Chakrabarty, Saakyan and Muresan (2021) advanced multilingual poetry translation by fine-tuning the mBART model on poetic texts, improving the preservation of both semantic meaning and stylistic fidelity across languages.

2.5.4 Advanced Techniques and Architectures

Innovations in model architectures have introduced new possibilities for poetry generation. Hu et al. (2024) explored Diffusion-based approach, proposing *PoetryDiffusion* models, that separate semantic content generation from metrical manipulation, allowing for joint control over meaning and poetic structure.

Pardinas et al. (2023) have incorporated techniques like Reinforcement Learning from Human Feedback (RLHF) to improve the quality of generated poetry, particularly in aspects like structure and language use. Zhang and Eger (2024) explored collaborative and multi-agent systems to increase diversity and novelty in generated poetry.

2.5.5 Human-AI Collaboration

Several studies have focused on creating systems that facilitate collaboration between humans and AI in the poetry writing process. These systems aim to enhance human creativity by providing suggestions, refining user inputs, or co-creating poems.

Interactive Poetry Generation Systems

Uthus, Voitovich and Mical (2022) proposed *Verse by Verse*, an interactive system that suggests verses in the style of classic American poets, allowing users to compose poems while maintaining control over the creative process. Similarly, Chakrabarty, Padmakumar and He (2022) introduced *CoPoet*, a system leveraging instruction-tuned large language models to assist users in writing poems based on specific requests. Their study demonstrated that users retain a significant portion of the AI's suggestions in their final compositions, highlighting the system's effectiveness in collaborative poetry creation.

Impact on Creativity and User Experience

Yang et al. (2022) explored the impact of human-AI collaboration in poetry writing on users' perceptions of creativity and workflow. In their interactive settings, the AI acted as an active writing partner, providing unexpected ideas that inspired and challenged human writers. However, Olatunji (2023) highlighted the need to ensure that such systems genuinely enhance human creativity rather than inadvertently diminishing it through over-reliance on AI-generated content.

2.6 Evaluation Methods

Evaluating AI-generated poetry is inherently complex due to the subjective nature of creativity and aesthetics. While virtually all research concerning poetry generation includes some form of evaluation to assess the quality of its outputs, this section specifically reviews studies where the primary focus is the development, analysis, or comparison of evaluation methods themselves, rather than the generation process. Researchers dedicated to this area have proposed various techniques and frameworks designed to measure diverse aspects of AI-generated verse, ranging from technical adherence to form to perceived creativity and emotional resonance.

2.6.1 Automated Evaluation

Automated metrics have become essential tools in evaluating poetry across various dimensions. For example, Agirrezabal, Oliveira and Ormazabal (2023) developed ERATO, a framework that analyzes technical aspects by examining poetic features such as rhyme schemes, meter, and structural adherence. Similarly, Chen et al. (2024) assessed lexical and semantic diversity to gauge the richness and originality of vocabulary and themes within poems. Furthermore, Boggia et al. (2022) employed language models to measure coherence and fluency, providing insights into the overall cohesiveness of poetic works. Sawicki et al. (2022) (Chapter 3) and Sawicki et al. (2023b) (Chapter 4) applied BERT and GPT-3 models fine-tuned as binary classifiers to evaluate the accuracy of style replication of GPT-generated poetry. Later on Sawicki et al. (2025) (Chapter 6) simulated the Consensual Assessment Technique (CAT) to perform "in-context" evaluations of poetry.

2.6.2 Human Ability to Differentiate Between AI and Human Poetry

Human judgment remains critical in evaluating poetry. Rahmeh (2023) conducted studies involving postgraduate English Literature students who read and rated poems on various dimensions, including aesthetic appreciation. Participants compared Shakespeare's "Sonnet 18" to an AI-generated sonnet on

enjoyment, emotional depth, and language complexity, using a 1-10 rating scale where higher scores indicated greater appreciation. The findings revealed that participants significantly favored Shakespeare's Sonnet 18 across all dimensions. For instance, Shakespeare's sonnet outscored the AI-generated poetry in enjoyment (mean scores of 8.08 vs. 6.10, respectively), emotional depth (7.94 vs. 5.80), and language complexity (8.10 vs. 6.55). While participants recognized AI's strengths in experimentation and efficiency, they identified substantial limitations in emotional resonance, human perspective, and thematic depth—suggesting that while AI poetry shows promise, it still falls short of capturing the full richness of human creative expression according to these raters.

Conversely, other research highlights the difficulty humans face in reliably distinguishing AI-generated text from human authorship, even when perceiving quality differences. For example, Gunser et al. (2022) investigated whether readers could differentiate between AI-based (GPT-2) and human-written continuations of classic literary texts. Their two studies found that participants generally exhibited low accuracy in correctly identifying the origin of the texts, misclassifying a significant portion of AI-generated continuations as human-written (e.g., 42.04% in Study 1) and vice-versa. Despite this difficulty in differentiation, participants were often overconfident in their judgments. Notably, even while struggling to identify authorship, participants consistently rated the AI-generated continuations lower on stylistic qualities such as being well-written, inspiring, fascinating, interesting, and aesthetic compared to both human-written and original classic continuations.

These contrasting findings underscore the complexity of human perception of AI-generated poetry: while qualitative differences may be perceived, leading to preferences for human work in some studies, this does not always translate into an ability to accurately determine a text's origin. This challenge in differentiation is further explored by studies such as Köbis and Mossink (2021) and Porter and Machery (2024), which also found that AI-generated poetry can be difficult to distinguish from human-written works, with participants sometimes even rating AI poems more favorably or misidentifying them as human-authored more often than chance.

2.6.3 Depth of Creativity and Emotional Resonance

Wöckener et al. (2021) highlighted the inherent subjectivity in poetry appreciation, noting the difficulty in standardizing evaluation criteria. They emphasized that capturing nuances such as emotional resonance, thematic depth, and stylistic authenticity requires sophisticated evaluation methods. Moreover, they stressed the importance of distinguishing between models that generate novel content and those that reproduce memorized text to assess genuine creativity.

Dai (2021) and Franceschelli and Musolesi (2024) report that while LLMs may capture the form of poetry, they often lack the ability to imbue it with genuine emotion or philosophical depth. They attributed

this limitation to the probabilistic nature of these models and their lack of consciousness, which can lead to outputs that feel superficial or lack authenticity. This perspective presents an interesting tension with the empirical studies discussed previously (Section 2.6.2) showing that human readers frequently cannot distinguish between AI-generated and human-written poetry in blind tests, suggesting that the perceived deficiencies in emotional resonance or depth may be more theoretical than experiential for many readers. Nevertheless, these scholars argue that the underlying generative mechanisms remain fundamentally different from human creative processes, even when the outputs appear indistinguishable.

2.6.4 Bias and Ethical Issues

Several researchers have highlighted ethical considerations in AI-generated poetry. Sheng and Uthus (2020) showed that models trained on large datasets may inadvertently perpetuate societal biases present in the data. The risks of deception posed by AI's ability to generate poetry indistinguishable from human work were discussed by Köbis and Mossink (2021) and D'Souza and Mimno (2023), noting potential situations where AI-generated content might be presented as human-authored without disclosure.

Additionally, Olatunji (2023) emphasized that bias in training data can lead to the exclusion of diverse voices and the perpetuation of stereotypes, necessitating deliberate efforts in data curation and model training. For example, Olatunji (2023) noted how certain social groups are underrepresented in widely used training datasets, while others are disproportionately represented, creating imbalances that reflect existing power structures.

2.7 Conclusion

This chapter has established the multidisciplinary foundations necessary for the research presented in this thesis. We have traced the technical evolution of Large Language Models, from the introduction of word embeddings to the transformative Transformer architecture and the GPT family, highlighting how these advancements have fundamentally altered the landscape of natural language generation. In parallel, we explored the theoretical dimensions of computational creativity, distinguishing between generative capability and intentional agency, and reviewing frameworks such as the Consensual Assessment Technique (CAT) which inform our evaluation strategies.

Our review of the history of digital poetry generation—from early rule-based systems to the pre-LLM neural era—reveals a trajectory toward greater linguistic fluency but highlights persistent challenges in achieving fine-grained stylistic control. As discussed in Section 2.5, the application of Large Language Models to this domain has yielded impressive advancements, enabling the creation of poetry that can, in some cases, rival human-authored works in form and aesthetic appeal. However, significant challenges

remain in capturing the depth of human creativity, emotion, intentionality, and style.

Addressing these challenges requires not only technical innovations but also thoughtful consideration of ethical implications and a commitment to preserving the richness of human artistic expression. Furthermore, the rapid progress in generation capabilities has created a critical bottleneck in evaluation, necessitating new methodologies that can keep pace with these models. These insights set the stage for Chapter 3, where we begin our practical investigation by exploring the capabilities and limitations of fine-tuning GPT-2 for stylistic replication.

Chapter 3

Training GPT-2 for Romantic-Era Poetry Generation

The primary aim of this chapter is to investigate the capacity of GPT-2 models to generate poetry in the specific styles of Lord Byron (1788-1824) and Percy Bysshe Shelley (1792-1822). Conducted in 2021, when fine-tuning GPT-2 represented the most viable approach for specialized text generation, this work explores the capabilities and limitations of these models when trained on limited domain-specific data.

A primary objective of this chapter is to address critical challenges in fine-tuning GPT-2 models for this stylistic poetry generation. This involves a detailed analysis of overfitting and its consequences, alongside the development of evaluation methodologies capable of assessing the authenticity and quality of generated poems beyond standard loss metrics.

The key contributions stemming from this investigation are:

1. A systematic **analysis of overfitting's impact** during GPT-2 fine-tuning, highlighting the necessity for multi-faceted evaluation beyond simple loss metrics.
2. The introduction of a **novel evaluation methodology** using BERT-based binary classifiers to assess stylistic authenticity in groups of generated poems.

To address these objectives, the chapter first defines the specific aspects of authorial style targeted in this study (Section 3.1) and details the experimental setup and data preparation (Sections 3.2 and 3.3). It then presents experiments evaluating overfitting through visual inspection and BLEU scores used unconventionally to detect repetition (Section 3.4). Subsequently, it details the use of BERT-based classifiers to assess stylistic fidelity (Section 3.5). The chapter concludes with a combined Discussion and Conclusion section (Section 3.7) that contextualizes the chapter's findings within computational creativity,

reflects on their broader relevance in light of subsequent LLM advancements, and discusses the limitations that motivate the research in the following chapter.

Research for this chapter was published in Sawicki et al. (2022).

3.1 Defining Authorial Style in Computational Generation

To evaluate whether GPT-2 can successfully simulate the voices of Byron and Shelley, we apply the operational definition of style established in the Introduction (Section 1.3). While we understand style fundamentally as a probabilistic shift in the model’s linguistic distribution, detecting that shift requires us to identify specific, observable markers relevant to the Romantic era. For the experiments in this chapter, we translate that high-level definition into four concrete layers of stylistic identity which the model must replicate:

- **Vocabulary and Tone:** The specific lexicon that anchors the text in the early 19th century. This includes the correct usage of archaic pronouns (“thou,” “thee”) and the specific contractions characteristic of the period’s poetic register (“’tis,” “o’er,” “e’er”).
- **Sentence Structure and Punctuation:** The syntactical habits that define the author’s rhythm. For Lord Byron, this specifically involves the frequent use of dashes (–) to create dramatic pauses or a conversational, interruptive flow, a signature we explicitly look for in the generated output.
- **Rhythm and Structure:** Although GPT-2 processes text as tokens rather than phonetic sounds, it must learn the visual and structural patterns of verse. We evaluate its ability to organize lines into coherent stanzas, maintain consistent line lengths, and attempt the rhyming schemes typical of the target authors.
- **Themes:** The semantic domain of the author. For Shelley, the model should gravitate towards nature, philosophy, and the sublime. For Byron, we look for satire, romantic irony, and the weary introspection of the Byronic hero.

A successful result in this chapter is defined by the model’s ability to generate text that captures these layers well enough that an automated classifier—trained to recognize these specific features—can correctly identify the output as belonging to the target author’s style.

3.2 GPT-2 Models and Experimental Setup

The experiments detailed in this chapter utilize variants of the GPT-2 model (Radford et al. 2019). As established in Chapter 2 (specifically Section 2.2.5), GPT models are based on the decoder component of

the Transformer architecture (Vaswani et al. 2017). During the initial phase of this research (2021-2022), fine-tuning publicly available GPT-2 models represented the most accessible approach for developing specialized generative capabilities, as fine-tuning for larger models like GPT-3 was not available.

A key aspect of this chapter involves comparing different implementations of GPT-2 available then. We utilized both the original implementation released by OpenAI (referred to herein as “Regular” models) (Shepperd, N. 2021) and versions available through the Hugging Face Transformers library (Wolf et al. 2019) (referred to herein as “LMH” for Language Modeling Head models). These implementations differ in their training interfaces and potentially performance characteristics, as explored later in this chapter. Due to hardware constraints, our experiments focused on the “Small” (124M parameters) and “Medium” (345M parameters) versions of both Regular and LMH implementations.

Regarding the generation process itself, we adhered to standard practices at the time. Unless otherwise specified, samples were generated using a default *temperature* of 1, a *top-k* setting of 50, and a *top-p* setting of 1 (adjusted from the LMH default of 0.9 for consistency, as noted in the original implementation details). It is important to acknowledge that fixing the generation parameters—specifically *temperature*, *top-k*, and *top-p*—represents an experimental limitation regarding the coverage of the models’ generative capabilities. In language modelling, these hyperparameters collectively govern the diversity and quality of the sampling strategy; systematically varying them could have yielded a broader spectrum of outputs and potentially mitigated degeneration issues like looping. However, conducting a comprehensive hyperparameter search to optimize these values was not feasible within the constraints of this study. Given the already extensive experimental matrix—comprising four distinct model architectures, two authorial datasets, and twenty-five fine-tuning checkpoints per run—introducing variable generation parameters would have resulted in a combinatorial explosion of outputs. This would have rendered the subsequent systematic evaluation, particularly the visual inspection of text quality, practically impossible. Therefore, fixed default parameters were utilized to maintain a manageable scope while ensuring a consistent baseline for comparison.

Furthermore, regarding the input provided to the models, we employed a strategy of “minimalistic prompting.” In this study, the objective was to evaluate the models’ intrinsic ability to replicate style via fine-tuning, rather than to control content via prompt engineering. Consequently, we sought to minimize the influence of the input text on the generated output. However, as the GPT-2 architecture requires an initial context to begin autoregressive generation, we selected short, thematically consistent anchor phrases to serve as seeds. For the Byron models, the prompt “The beating heart” was utilized, and for the Shelley models, “The eternal sky”. These prompts were chosen to set an appropriate poetic tone while remaining sufficiently open-ended to grant the models maximum freedom in determining the subsequent content and structure.

Further details specific to the experimental setups, such as the duration of fine-tuning runs, the number

Author	Org. Size	Org. Length (lines)	Final Size	Final Length (lines)
Byron	7.2 MB	183643	2.4 MB	62947
Shelley	2.3 MB	59207	0.98 MB	29151

Table 1: Statistical summary of the datasets used for fine-tuning. The table compares the original collected works (Org.) against the final pre-processed versions, detailing file sizes (in MB) and total line counts.

of samples generated per checkpoint, and the token length of generated samples, are described in the relevant setup descriptions (Sections 3.3.2 and 3.3.3).

3.3 Data preparation

3.3.1 Original datasets

Our main interest is the generation and evaluation of poetry in the style of a specific author. The Romantic era was selected due to the distinct and often formal stylistic features prevalent among its prominent poets, offering a clear target for stylistic replication with the relatively less powerful GPT-2 models. Furthermore, extensive corpora for poets such as Lord Byron (1788-1824) and Percy Bysshe Shelley (1792-1822) are readily available in the public domain (e.g., from Project Gutenberg), facilitating data collection for fine-tuning.

The datasets for Byron and Shelley were created from their collected works, downloaded from Gutenberg.org (Project Gutenberg 2024). This involved removing all introductions, forewords, footnotes, and generic Gutenberg text at the end of the file. All extended Latin characters were replaced with their closest matching ASCII character (for example, “ \tilde{A} ” is replaced with “A”, “ \tilde{a} ” with “a”, etc.), and sequences of multiple blank lines were condensed to a single blank line. We also removed metadata such as page numbers and end-of-line verse numbers. However, poems’ titles and chapter numbers were retained as they contribute to the author’s “style” and help preserve the separation of individual poems. Being aware of the destructive impact that artifacts of poor data pre-processing can have on GPT-2 output, particular attention was paid to this task.

Additionally, we have removed all plays, thus leaving only poetic works. The purpose of this pre-processing was to leave only the poetic text written by the authors themselves. For the subsequent fine-tuning and evaluation tasks described in this chapter, these pre-processed datasets were divided using a standard 70/30 train/evaluation split.

3.3.2 Setup 1 for visual and BLEU evaluations

We fine-tune all four GPT-2 models used in this study (Regular Small, LMH Small, Regular Medium, LMH Medium) on both datasets (Byron and Shelley). For the visual and BLEU evaluation in Experiments 1 and 2, we fine-tune the GPT-2 models for 250K steps, generating 100 samples at each 10K steps interval. The samples are generated with a length of 1000 tokens (the maximum sample length for these models is 1024 tokens). We generate only 100 samples at each checkpoint because of the time it takes to generate full-length samples (for example the LMH Medium model running on Nvidia P100 GPU takes around 2 minutes per sample). Thus we obtain 8 sets of 2500 samples, four for each author. Figure 11 shows the training loss and evaluation loss for the Regular Small and LMH Small models fine-tuned on the Byron dataset for 250K steps (results for medium models and for models fine-tuned on the Shelley dataset are very similar, and therefore are not presented here). We can see that the lowest evaluation loss is achieved very early in the fine-tuning process: for LMH models, this occurs around 5000 fine-tuning steps, and for the Regular models even sooner, around 1700 steps.

We believe this is because the datasets used for fine-tuning are relatively small, and the models become overfitted fairly early. In this study, we investigate whether the point of the lowest evaluation loss is optimal for early stopping of the fine-tuning process, and to get deeper insights into the behaviour of the GPT-2 models, we evaluate the actual quality of the generated samples. To that end, we conduct a number of evaluations of samples generated at specific checkpoints. This will be further described in the later sections.

3.3.3 Setup 2 for BERT evaluations

Datasets for the visual and BLEU evaluations created in Setup 1 have an insufficient number of samples per checkpoint (only 100) to be used for training the BERT-based classifiers. For this reason, we create a separate set of 8 datasets by fine-tuning all of our GPT-2 models on both datasets for 10K steps and generate 1K samples at each 1K steps checkpoint. We have chosen this span of checkpoints because it covers the sweet spot where the evaluation error is at the lowest, and we can observe the quality of the samples immediately before and after that point. Thus we obtain 8 datasets of 10K samples. The samples are limited to 600 tokens, since as we explain later, we use only the first 20 lines of each sample in Experiments 3 and 4.

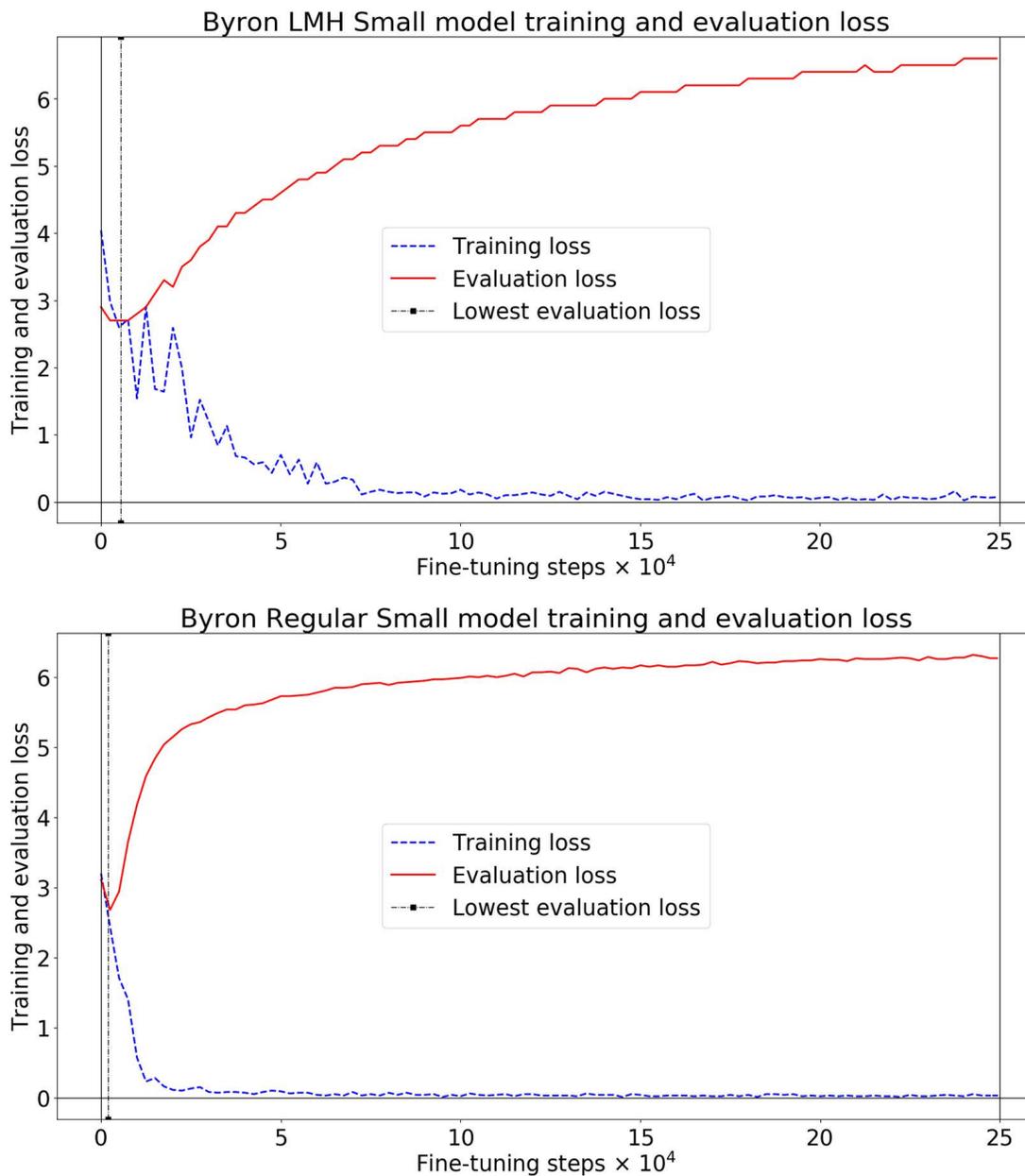


Figure 11: Training and evaluation loss for the LMH Small model (Top) and the Regular Small model (Bottom) fine-tuned for 250K steps on the dataset of Collected Works of Lord Byron.

3.4 Part 1—Evaluation of the overfitted models

In this section, we analyse the impact of overfitting on the quality of the produced samples. This is to emphasize the importance of early stopping of the fine-tuning process and to explore how the existing quantitative metrics (such as BLEU) correlate with the overfitting of GPT-2.

3.4.1 Qualitative Overview of Generated Output

Before analyzing the statistical evidence of overfitting, it is necessary to examine the qualitative nature of the text generated by the various GPT-2 models. While the models often struggled to maintain long-range coherence over the full 1024-token context window, they were capable of producing linguistically fluent fragments that mimicked the target authors' styles.

To illustrate the capabilities of the different architectures, we present samples extracted from the continuous generation streams of the Regular Medium and LMH Medium model variants.

Stylistic Fragments: Lord Byron. The following fragments demonstrate the models' attempts to capture aspects of Byron's style:

Regular Medium Model

Hence from my native land—lost without trace
 Of hills, or wilderness scarce discernible,
 Here Adeline descended with her brother;
 The deadly wind and damp torrent bore
 His motionless bark, unheeded by Heaven:
 Through thickets and crevices on the shore
 The dim mists seemed little changed by the spray;
 The quitting star was little screened by the cloud,
 Nor the rising pang by the approaching night!

LMH Medium Model

The storm, and the desert, and the sky,
 The desert—the storm—and the mountains afar!
 The land—the land—the ocean—there is no rest;
 No land, no earth, no sea, no home;
 No home, no grave, no sepulchre—
 No tomb, no tombstone, save one—

Save one—the very grave that knows
The name of me.

Stylistic Fragments: P.B. Shelley. The following fragments demonstrate the models' attempts to capture aspects of Shelley's style:

Regular Medium Model

It is not blasphemy to hope that Heaven
More perfectly will give those nameless joys
Which throb within the pulses of the blood
And sweeten all that bitterness which Earth
Infuses in the heaven-born soul. O thou,
Whose dear love gleamed upon the gloomy path
Which this lone spirit travelled, drear and cold,
Now swiftly lead me to those awful limits
Which mark the bounds of Time and of the Space
Where Time shall be no more.

LMH Medium Model

This is what Pleasure is: she who gave
The spark that blazed, and from its flame
Flashed all kindly light, and Heaven's ebon flame
To decorate the naked earth; she who drew
The breath of life into all things,
And breathed it thence into the eternal hills;
And from her being, ever new and bright,
Burns a warmer tide of glory, which fills
The habitable earth with her sweet presence.

Complete Generated Samples. The following examples display the full output generated by the GPT-2 Medium Regular models for both target authors, presented here for illustrative purposes:

Complete Sample in the style of Lord Byron (GPT-2 Medium "Regular"):

"The beating heart!"

..... POETIC BLISTER.

Or, to put the case, a European
The same ancestor boasted:
He was not the monarch of British soil,
Nor the son of the phalanx of Holland;
He could cite an escaped slave and allow
His pardoning right to adulterers;
The Britons shall deem him a barbarian,
Who draws verses from the lips of his fathers!

SOUNDS OF LIFE.

Born near Lochra, a remote bight,
Myself peculiarly trained in the breed,
To the vague flickering melody there is,
Which roars from the farthest moment;
That pensive, remote exhalation,
Of hearts already boisterous from childhood:
The one winds at midnight, the other has
A choral company in Sweden,
Lulling, completely exhausted, winds for a league.

I.

Their solemn rendezvous, and weary sighing,
Said “gone forever far away,” because nothing
Went far away—and
Thus, as if waiting for an advent,
In fragments, I fan the reposing talent.
In this I sped along the road, drunken with citizenship, gentle companions,
And sister tongues, but scarce understood.

II.

Their hounds, clumsy and shadowy, sickened and wailed;
And my dear Eunice thawed to join
In the set hinds howling with rapture,
For thrice, thrice her Maker has said—
And we must remember now—this quiet, sublime buzz
Of purer joy could charm a score of hearts.

III.

They wagged their high, fawn-like wings, for instance,
With, “My sweet Eunice—or as yet to be throned over!”
And such were their actions as were meant:
The heart swelling with the determining extent, made
At last, a storm, which broke on the sensitive point.
Grace! The heart may swell from its due memory,
But swell not from love, and not from good or ill,
Until it withdraws almost from the breast:
The Fremont was angry, and shrieked vehemence
Not from an affectionate mother, but a strife.

BOTH.

“Her heart is sufficiently girt for the battle
The blood of mummers is moist and flow....”
With health—full of self-pity, of ardour,
Euphoric orator, I gladly commend
To every lover or fairy an enflamed vein,
Which will never slacken, like some slumbering clay,
The frown of disinclination; for my years,
Have made my soul heavy and void of all shape,
Perhaps it will benefit its tender bosom from this strife.

TABLE.

He loved her, but hate was not scathed;
Even in pangs he sought the newly-fledged breach:
A still louder anger sufficed for rival
Praise, or esteem, but only more attended.

Complete Sample in the style of P.B. Shelley (GPT-2 Medium “Regular”):

The eternal sky
(The gods could not wash away; they had iron in their blood,
While still the terrible flood of ice had rolled down his strand;
Hydarn, the victim of Aiwass,)
Would sigh and wail and drink her last;
While for Phoebus and his news the wild stream
Swung ten thousand furlongs on the desolate plain,
Till it laboured and the waves threw down the banks;
And the rainy dale, and the dismal woods,
And the quivering dells of the forest’s wood,
With a faint earthquake would meet in silence,
Till with the loud crash of a thousand trumpets
The sun rose, sinking on the earth and Heaven.

DEATH.

(No sooner thus is Death gradually passed than all horrors
Turn of mind, shape, and motion into stillness,
Might stretch its arms around his motionless form.)

He does not wake, but he but dimly perceives
That sleep is dead, and that Death lies dreaming on
The last leafless disk of Heaven’s passionate moor,
On which the venomed snakes and the myriad flies
Burn with long life and hopes, and no ill,

No passion, but love and ecstasy through his brain;
There is no stranger's gauche look of pain
Struggling in the tearless lips where ONE shone fast,
Gulfing and mournfully failing morn,—'tis pity.

DEATH.

Lo! The breath of Death was moving on my hair,
And the cold leaves around me were withering,
And around my body, in ages of happy youth,
I now lay, and the sound of the dying breath
Was spreading and spreading around me; and I lost
The breath that pierced me, and the involuntary pulsation

Grew nearer and nearer, until at last it seemed to bleed,
And died on my dear breast, and painfully
I bit it, till my blood ran o'er my lips
And in those dusky eyes my fingers both drew
The breath that from its wound in my brain gave
My madness to live. Death hath killed the proud Ingenious,
The obstacle of truth's welcome affright,
The hammer of truth's progress. Death hath killed the Green Death,
The oak at my door, whose head has driven the seed

Of Anarchy away whence it springeth and fell. Death hath killed the stains
Of man's perpetual sin, all wound and wrenched
From innocent destruction, or danger pure;
Death's mummer, to whom I stood, and I heard
The downward breath of life, or spirit of thought,
Which soon preying itself on my solitary soul,
This phantasmal portrait of odium,
Grew on my being like a swarm of bees,

Which approach the flower; and after a while
Oh! Re-awaken my consciousness by your lips!
I grieved that hate of truth, or innocence, or fate,
Did reign amid Life's degenerate giant shade,
Mocking and blaspheming everything noble
Which mewed beneath the spreading beam of Time.
Death is Lorn! Embalm not, perhaps, blood,
Or reviving in thy potent morn, my spirit.

TO: CHAPTER V.

Poets weep, and Cloisters rise from the grave,
A morality from designs immortal,
What if pity deceived to cut flesh on blow?
The waters of unfettered profanation,

'ALMIGHTY DREAD.'

Hast thou in thy native island wereland,
Wasted not humour in such profanation?
He of whom Poesy the Deist readeth;
He who overthrows all former laws,
And builds Temples of death, and enacts
Robbery and fear:—our Eclectes
Signals to the King pledges of hatred;
His Priests swear to hang him at Midnight;
Treason or Death, if they see the flame
Blazing in his eyes, will kiss each other at Day.

THOUGHTS.

Holy thoughts,—And holy thoughts, how dost thou wane!
 The changes thou didst see in my breast
 Of misty tears, and sunless bliss, which soon
 Had perished—and thou didst fade away, leaving
 My conception to the ambition
 Of machines; yet still I conceived
 That thee, my own tender self, were my refuge
 To succour that which most ravished me, then
 O'er all things I dost follow the end.

3.4.2 Experiment 1—Visual evaluation of text quality

During this research, we have observed that while generating text with the full length possible, many samples come with significant errors. We have decided to establish whether there is any regularity in the production of the malformed samples. For this purpose, we analyse the datasets from Setup 1. From every 100 samples generated at a specific checkpoint, 10 samples are selected at random and evaluated manually, using the following procedure: when looking at the sample, we check at which point within the sample the text becomes unintelligible or contains obvious errors. We take note of the line number where this happens, and we take note of the total number of lines in the sample (including the blank lines at the end). Then we calculate the percentage of the correctly produced lines. After that, we calculate the average value for those 10 samples. We repeat this for all 25 checkpoints. The results for both datasets are shown in Figure 12.

We can see that the Regular models score almost 100% across the whole range of fine-tuning checkpoints. For the LMH models, the percentage of correctly generated text within a sample is at its best at 10K and 20K checkpoints, and after that, it rapidly decreases to around 35% for the remaining checkpoints.

We have observed that once the errors start appearing in the samples generated by the LMH models, the remainder of the sample is almost always malformed. In contrast, the Regular models occasionally produce a few malformed lines in the middle of the sample, but the subsequent text is consistent again. The LMH models' output does not have this “self-repairing” property. We are aware that these results could be different if much larger or much smaller datasets were used for fine-tuning. The reason we chose datasets of this size is because of our objective of style preservation of an individual poet.

A well-formed sample with a Byron-style text of 1000 tokens usually spans around 45 to 80 lines in

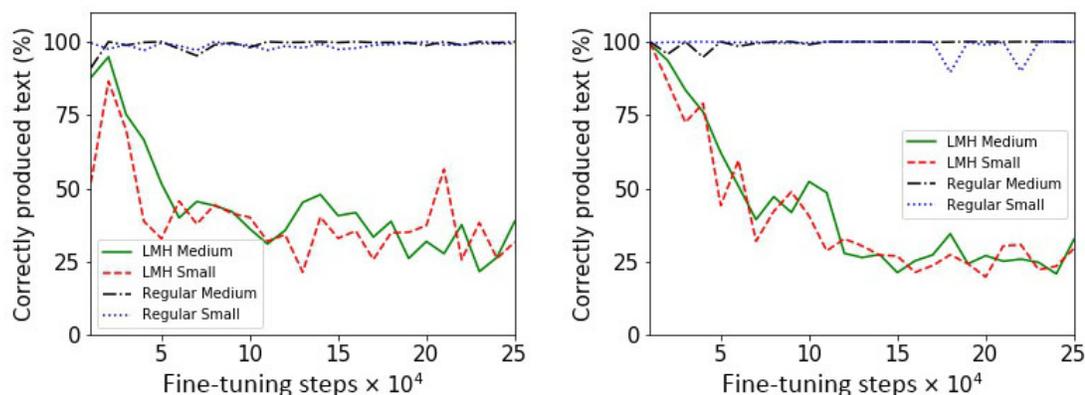


Figure 12: Results of the visual evaluation of text quality for samples generated from the Byron (left) and Shelley (right) datasets.

a text file. However, the malformed samples from the LMH models could sometimes exceed 180 lines, of which often only around 30 lines at the top of the sample are of good quality. With the Shelley-style samples, the malformed samples can exceed 250 lines, with more or less the same proportion of correctly produced text. This is because the number of tokens per line plummets, and many blank lines are inserted into the sample. This “stretching” phenomenon was not observed in the samples produced by the Regular models. One could raise a question whether these results were caused by poor data pre-processing. This, however, is unlikely since our data pre-processing was rigorous and comprehensive, as described in the Section 3.3 on Data Preparation. Given the quality and fineness of data used for fine-tuning, it is clear that the errors in the samples must be caused by the properties of the models themselves.

The following example illustrates the specific degradation issues observed in the LMH models. As the model overfits, the text loses semantic coherence, often degenerating into repetitive loops or the “stretching” phenomenon where lines become very short and are separated by excessive whitespace.

Malformed Byron Sample (LMH)

The beating heart , the sighing head,
 The quick successive throbs,
 Which fill
 The bursting thundering here,
 Which for a moment all was done,

 But the spell is broke
 In one eternal gush of blood,

And no second skin
Of human shape
Can rend the yet warm skin
Of that glowing heart
Which beat to its wonted fires
With one eternal flame,
Which, kindled by another,
In friendship and peace,
The twelfth heaven brought forth
The blackening fires above,
The lightning of the sky
Sent
A solitude
THE deadliest the Tempest of the skies.

II.

A SONG OF SAUL.

THE SONG OF SAUL.

WRITTEN BY TITLED BY

I.

SONG.

I.

I SONG.

I WEEPING.

CANTO LOVE.

FRAGMENT.

FROM LINES.

I.

JOHN KEATS.

DRUNKS A BEAUTY.

I SONG.

FROM THE SONG.

That rocked the deep billows

Ocean's foam,

And I loved,

To him dead,

Muse

unk OR playful;

Which loved,

LINES.

WHEN HIS smile;

That loved,

ADDRESS

LINES.

A WELLENES

LINES.

OR sung

WRITTENESS.

DRUNGMENTS

Because of the malformed text in long samples, in the subsequent experiments with the BERT-based classifiers, we will limit the sample length to the first 20 lines of text from each sample. This is because our goal in the second part of this study is to evaluate only well-formed outputs, instead of learning to spot obvious errors, like repeated lines, garbled and unintelligible text, etc.

The results of the visual evaluation in Figure 12 show a deficiency of the LMH models after 10K-20K training steps with these specific datasets. When these results are contrasted with Figure 11, one can notice a clear correlation between overfitting—quantified by a high evaluation loss in Figure 11—and the ratio of malformed lines in Figure 12. Such a correlation is not present in the results of the Regular models, however, and their results in Figure 12 do not detect any malformed text according to the visual evaluation. For this reason, we perform a BLEU evaluation in the next experiment to see if the effect of overfitting can be uncovered in the samples from the Regular models.

3.4.3 Experiment 2—What does the BLEU evaluation tell us?

The visual evaluation has informed us that samples from the Regular models appear to be correctly produced across all the checkpoints from 10K to 250K fine-tuning steps, regardless of the increasing evaluation loss between those checkpoints. But, are there any noticeable and measurable changes in text quality between those checkpoints? In order to establish that, we perform a BLEU evaluation of those samples against the original dataset.

Bilingual Evaluation Understudy (BLEU) was originally designed to evaluate machine translation (Papineni et al. 2002), where we have the master translation of the sentence in the source language, and a candidate translation produced by the system. BLEU compares and counts the matching n-grams between the master and the candidate, resulting in a score between 0 and 1. The closer the score is to 1, the better the translation, because this indicates that the candidate translation is more similar to the master translation. While not designed for this purpose, BLEU is sometimes used to evaluate text quality (Yu et al. 2017), but when used for this purpose, it suffers from several deficiencies. First, BLEU cannot capture semantic similarity beyond n-gram matching, failing to recognize paraphrases or synonyms that

maintain meaning (Reiter 2018). Second, it shows poor correlation with human judgments of text quality and fluency (Novikova et al. 2017; Callison-Burch, Osborne and Koehn 2006). Third, BLEU penalizes any deviation from reference texts, even when such deviations might represent creative or stylistically superior writing (Liu et al. 2016). Finally, as BLEU was designed for comparing translations against reference texts, it struggles in generative contexts where multiple diverse outputs could be equally valid (Mathur, Baldwin and Cohn 2020).

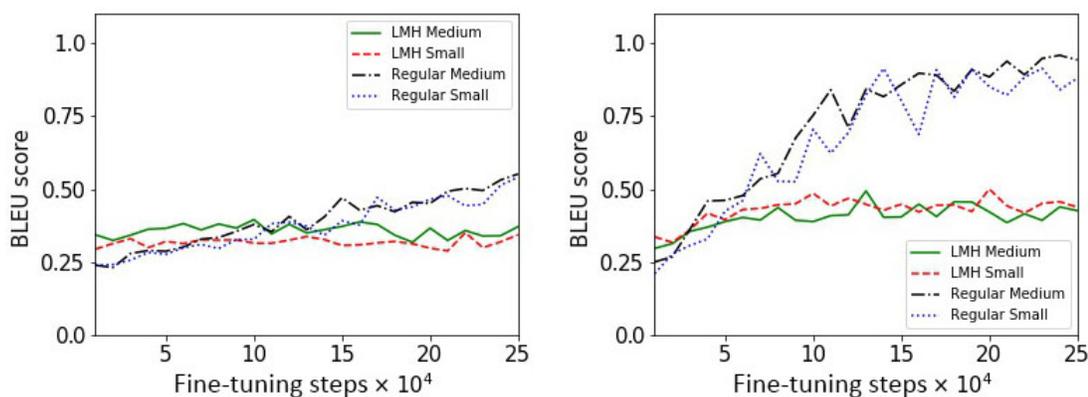


Figure 13: BLEU scores for Byron (left) and Shelley (right) calculated for samples with 1000 tokens length.

Our objective, however, is to measure only the n-gram based similarity between the samples and the original text. We therefore expect that BLEU is an appropriate algorithm for our application because we have two types of text to compare, albeit we interpret the scores differently. Unlike in the translation tasks, in our research, we are aiming at a lower score, which indicates that fewer n-grams in the sample were copied from the original dataset, thus demonstrating a higher level of originality. In other words, we treat the BLEU score as a plagiarism detector on the n-gram level, which would be quantified by a high score. We use it to explore if there are any trends in the averaged BLEU scores between consecutive checkpoints. In our application of BLEU, we score each sample against the original dataset, and then average the results, similarly to what was proposed by Yu et al. (2017) and implemented by Wesek (2021). The implementation of BLEU used in our code was from the NLTK library (NLTK Project 2021).

Like in Experiment 1, we follow Setup 1, and we use 100 samples for each checkpoint from 10K to 250K. Samples have the length of 1000 tokens. We compute BLEU for all the 100 samples at each specific checkpoint, and we take the mean of those values to obtain a single value per checkpoint. Figure 13 shows that the BLEU score consistently increases with the model fine-tuning steps for both Regular models. This indicates an increasing similarity of the samples compared to the original dataset, when the models are fine-tuned for longer. This increasing BLEU score basically means that GPT-2 plagiarizes n-grams

from the original dataset. On the other hand, the BLEU scores for samples from the LMH models do not increase in the same way. This is because of the increasingly high amount of malformed text, which prevents the BLEU score from rising. When we evaluated the samples truncated to 200 tokens (which discarded all the malformed text in the samples from the LMH models), the increase of the BLEU scores for both types of models was very similar. In other words, they consistently rose with the number of fine-tuning steps.

Altogether, we have observed that overfitting in the LMH models is easy to spot because the samples are malformed in an obvious way, but we must remain cautious about overfitting the Regular models, where the results of overfitting are not noticeable by our visual evaluation (Figure 12). Furthermore, the samples from the Regular models appear to be well-formed across all the checkpoints, while containing increasingly higher levels of plagiarized text or n-grams. This means that both automated and human evaluators could be misled by such outputs, since we cannot expect them to have memorized the original text. According to the results of this experiment, it is advisable to stop the fine-tuning process when the evaluation loss is minimised or soon after when the samples start to be well produced after the initial learning of the model. Excessive training can lead to plagiarised n-grams, with even entire blocks of text repeated in GPT-2’s output for the numbers of fine-tuning steps above 100K on our datasets.

To illustrate this phenomenon, the following sample generated by the Regular model after extensive fine-tuning (100K steps) demonstrate how the model reproduces large segments of the training data. Unlike the malformed output of the LMH model, these sample remains structurally coherent but heavily plagiarizes the original training corpora.

Plagiarized Sample (Regular Model - Shelley)

Text in **bold** was copied verbatim from “Ode to Liberty” and “Revolt of Islam, Canto VIII” by Percy Bysshe Shelley.

The eternal sky
 ’Tis **Disney** there, with his best and brightest,
 To show what I mean by fairies,
 And how they like to be unseen.

17.

This was the tenour of my waking life
 Like a ocean, whose waves are ever screaming,
 Till an island in the shower of the Sun,
 That is Pakku’s coathee, or Sumner’s Ithaca;–

A loose cannon, a rogue in the first flush
Of playing defense, a hater of virtue high,
A thief, who steals what he feels, and gives
Sword to blood-encircled thieves,
And they curse their repose with five languages persecuted
Around them,—d disdain for ever to speak again.

18.

”Disdain not thou, at thine appointed term,
To set thine armed heel on this reluctant worm
That crawls within his lattice cave,
His mature decision receive thee back;
And if any soft passion thou lovest
Should make thee despoil it, cajole thou,
And use thou this mighty Serpent to thine work?

19.

”Yes, I would not have thee transferred
But that thou shouldst cease to be such a bitch,
And that the forces of life should no more vex
Thy darkness—that the proud mind of man may know
How far thou canst fly, forever
Those woe-stricken slaves with one word destroyed
My empire, and taughtiquet the nations
How to build thy pure sanctuary there.

20.

”Surround me now, please—ruling the ball
Round me round my hungrily-piled palaces,
My millions trod—my soul was grown
Through torture and terror; a vile sting
Crept with imperishable germs of ill,
Till human thoughts might kneel alone,
Each before the judgement-throne

**Of its own aweless soul, or of the Power unknown!
 Oh, that the words which make the thoughts obscure
 From which they spring, as clouds of glimmering dew
 From a white lake blot Heaven's blue portraiture,
 Were stripped of their thin masks and various hue
 And frowns and smiles and splendours not their own,
 Till in the nakedness of false and true
 They stand before their Lord, each to receive its due!**

21.

'Ordain, oh, English people, this is what I preach!
 Come near my grave, and quench thy thirst
 To the bloom of its neglected fruit,
 Which - thou blothest from this shore
 Of life's blue and final stream—
 Thou wilt leave unfelt, and cause me thine
 To flourish in this faded day, and blot thee
 From my memory, alas!
 For after thy ruin, man must leave
 This earth a nobler being, more free,
 Than he who created thee in his mind.

22.

**'O Love, who to the hearts of wandering men
 Art as the calm to Ocean's weary waves!
 Justice, or Truth, or Joy! those only can
 From slavery and religion's labyrinth caves
 Guide us, as one clear star the seaman saves.
 To give to all an equal share of good,
 To track the steps of Freedom, though through graves
 Her dim footsteps lie, to suffer all in patient mood,
 To weep for crime, though stained with thy friend's dearest blood,—**

23.

**'To feel the peace of self-contentment's lot,
 To own all sympathies, and outrage none,
 And in the inmost bowers of sense and thought,
 Until life's sunny day is quite gone down,
 To sit and smile with Joy, or, not alone,
 To kiss salt tears from the worn cheek of Woe;
 To live, as if to love and live were one,—
 This is not faith or law, nor those who bow
 To thrones on Heaven or Earth, such destiny may know.**

24.

**'But children near their parents tremble now,
 Because they must obey—one rules another,
 And as one Power rules both high and low,
 So man is made the captive of his brother,
 And Hate is throned on high with Fear her mother,
 Above the Highest—and those fountain-cells,
 Whence love yet flowed when faith had choked all other,
 Are darkened—Woman as the bond-slave dwells
 Of man, a slave; and life is poisoned in its wells.**

25.

'Man seeks for gold in mines, that he may weave

A

This finding has significant methodological implications for computational creativity research. When a language model reproduces extensive segments from its training data rather than generating novel content in the author's style, it fundamentally fails as a creative system. This represents a significant evaluation challenge: without explicitly checking for repeated content against the original corpus, researchers might incorrectly attribute creative capability to what is essentially a sophisticated text memorization system. Moreover, in practical applications, such plagiarized outputs could potentially raise copyright and ethical concerns if presented as original AI-generated content. This underscores the importance of implementing robust plagiarism detection as a standard evaluation metric in generative AI systems.

A base requirement for GPT-2 to be creative is that it creates semantically and syntactically correct output. However, one way for it to do so is to just copy the source material, so high BLEU scores (or

another measure for plagiarism detection) can indicate that the system is moving away from novelty, which is also fundamentally necessary for creativity. This tension highlights a central goal in computational poetry generation: to produce text that appropriately balances adherence to learned stylistic patterns with original, novel content. Our ultimate aim is not merely to replicate existing poetry but to develop systems capable of generating text that captures an author’s essence while contributing something new. As such, using a system like BLEU can be helpful in experiments with GPT-2-based poetry to establish when the model shifts from generative creativity to mere reproduction of training data.

3.5 Part 2—BERT evaluation of produced samples

In this section, we perform two experiments using BERT-based binary classifiers to establish which of the four GPT-2 models (Regular Small, Regular Medium, LMH Small, and LMH Medium) is most effective at replicating the distinctive styles of Byron and Shelley. While our previous evaluations using visual inspection and BLEU scores provided insights into text quality and potential plagiarism, they don’t directly measure how well each model captures the nuanced stylistic elements that characterize these Romantic poets’ works.

BERT (Bidirectional Encoder Representations from Transformers) offers a powerful approach for this evaluation because it has been pre-trained on large text corpora and can be fine-tuned to recognize subtle patterns in text. We leverage this capability through two complementary experimental setups:

1. **Experiment 3:** A “Turing test” approach where we fine-tune BERT to distinguish between original poems and GPT-2 generated samples. If GPT-2 successfully captures the author’s style, the BERT classifier should perform close to random (50% accuracy), indicating it cannot reliably tell the difference between real and generated text.
2. **Experiment 4:** A novel comparative approach where we train BERT to distinguish between texts generated by two different GPT-2 models, then test it on the original poet’s work. This allows us to directly compare which one of the models produces output that more closely resembles the original author’s style.

Both experiments focus on the evaluation of correctly formed text only, using samples generated before significant plagiarism occurs (at 1K-10K fine-tuning steps). We analyze only the first 20 lines of each sample to ensure that only the correctly produced text is analyzed.

3.5.1 Experiment 3—Can fine-tuned GPT-2 outwit BERT?

This experiment aims to determine if GPT-2’s outputs can achieve a quality high enough to confuse a BERT-based classifier trained to distinguish between generated text and the original poet’s work.

Building on findings from Section 3.4, which highlighted issues with plagiarism in later fine-tuning stages (especially for Regular models, see Section 3.4.3) and malformed text in LMH model samples (see Section 3.4.2), we implement specific data selection strategies for this experiment:

- Samples are sourced from early fine-tuning checkpoints (1K to 10K steps) to minimize the impact of plagiarism.
- Only the first 20 lines of each sample (both generated and original) are used, ensuring that BERT evaluates correctly formed text, not obvious errors.

For each BERT-based binary classifier, a balanced dataset of 2K samples (1K per label: 0 for original author, 1 for GPT-2 generated) was prepared. This dataset was split into 70% training, 25% testing, and 5% validation. We utilized the “bert-base-uncased” model from the Transformers library, fine-tuned for 20 epochs using the Adam optimizer with a learning rate of $2e-5$. Classification accuracy on the test data serves as the primary evaluation metric, suitable for this balanced problem.

As detailed in Setup 2 (Section 3.3.3), GPT-2 models were fine-tuned for 10K steps, with 1K samples generated at each 1K-step interval. This checkpoint range covers the period of optimal evaluation loss (as seen in Figure 11) and allows for the training of ten distinct BERT classifiers per dataset/GPT-2 model pair, enabling analysis across these early fine-tuning stages. Results of these evaluations are presented in Table 14.

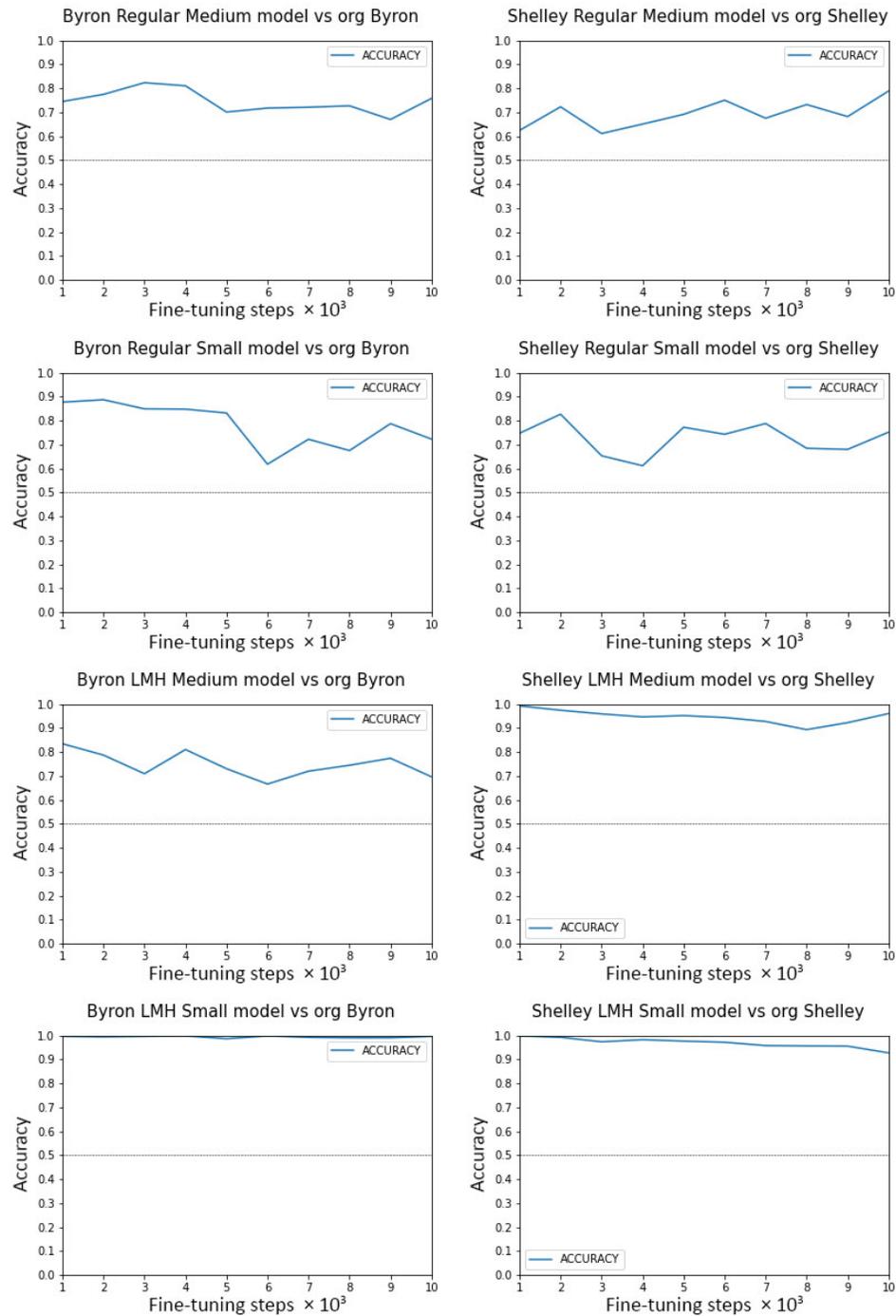


Figure 14: Results of the BERT evaluation of the Byron (left) and Shelley (right) checkpoint samples produced by the four different GPT-2 models. Row 1 (top): Regular Medium models, Row 2: Regular Small, Row 3: LMH Medium, Row 4 (bottom): LMH Small.

It is important to clarify the interpretation of results in this context. While in most classification tasks, accuracy values approaching 1.0 signify desirable performance, our interpretation here is different. In this experiment, GPT-2’s output is considered indistinguishable from the original author’s text if the BERT-based classifier performs at chance level, akin to a random classifier. Given that this is a balanced two-class problem, a random classifier would have an expected accuracy of 0.5. Conversely, accuracies approaching 0.0 or 1.0 would indicate that the classifier can still effectively distinguish between the two classes. Therefore, an accuracy of approximately 0.5 is the desirable target, serving as evidence that GPT-2 has successfully replicated the author’s style to a degree that confuses the classifier. This can be conceptualized as a form of “Turing test,” where success is marked by the BERT classifier’s inability to differentiate reliably. Such adversarial evaluation approaches have precedents in NLP, employing both human (Köbis and Mossink 2021) and automated evaluators (Bowman et al. 2015).

Figure 14 shows the classification results for Byron and Shelley’s datasets generated from all four GPT-2 models used in this study. The results show that the Regular models perform well on all checkpoints on both datasets, but interestingly the Regular Small model required 6K steps to reach its optimal performance (on Byron dataset), while its lowest evaluation loss is at 1700 steps (Figure 11). This indicates that we cannot rely on the evaluation loss alone, but instead, we may want to analyze the models’ output to establish the optimal early stopping time. The LMH Medium model performs well on Byron, but very poorly on Shelley. The LMH Small models have the lowest scores on both datasets. Thus, the Regular models appear to be a more reliable choice.

All models appear to have similar, stable performance across all 10 checkpoints (Figure 14), and thus these results do not correlate with the evaluation loss. This is because they are for the numbers of fine-tuning steps when no strong overfitting is observed (Figure 11) and the BLEU scores did not increase significantly yet showing over-training (Figure 13).

In the next experiment, we use BERT and Setup 2 again to compare the GPT-2 implementations, but using a different experimental design.

3.5.2 Experiment 4—Which GPT-2 is better at replicating the author?

In the previous experiment, we were classifying samples from the original dataset (label 0) against samples from a specific GPT-2 model (label 1). This gave us some indication as to which model is better at replicating the authors’ style.

Here, we propose a novel setup for text evaluation with the BERT-based classifier. This time we take samples from two different GPT-2 models, which we assign labels 0 and 1, and we classify against them only the samples from the original author’s writing. The accuracy is averaged, and it indicates to which models’ output the samples from the original author are closer.

To train the classifier, we use the dataset of 1K samples generated from two different GPT-2 models in a given pair after 10K steps of training. We selected this number of fine-tuning steps because (according to our previous experiments) both the evaluation loss (Figure 11) and the BLEU scores (Figure 13) show that we are not comparing overfitted models that plagiarize the original works. As explained before, only the first 20 lines of each sample are used in this experiment (see Setup 2). Just like in the previous experiment, we use “bert-base-uncased” from the Transformers library, which is trained for 20 epochs, with the Adam optimizer, and a learning rate of $2e-5$. Every classifier is tested on an additional test dataset of 1K samples randomly selected from the original authors’ corpus, each sample 20 lines in length. The results are averaged, giving a single value of accuracy. This value indicates which label the original dataset is closer to, i.e., which GPT-2 generates text more similar to the original work. Since we have four different models, we can create six possible pairs (Table 2) for each dataset.

Label 0	Score	Label 1
<i>Byron’s Poetry</i>		
Regular Medium	0.32	<i>LMH Medium</i>
Regular Small	0.08	<i>LMH Small</i>
Regular Medium	0.32	<i>Regular Small</i>
LMH Medium	0.09	<i>LMH Small</i>
<i>Regular Small</i>	0.51	LMH Medium
Regular Medium	0.08	<i>LMH Small</i>
<i>Shelley’s Poetry</i>		
Regular Medium	0.31	<i>LMH Medium</i>
Regular Small	0.49	<i>LMH Small</i>
Regular Medium	0.28	<i>Regular Small</i>
<i>LMH Medium</i>	0.54	LMH Small
Regular Small	0.49	<i>LMH Medium</i>
Regular Medium	0.31	<i>LMH Small</i>

Table 2: Results of Experiment 4: BERT classification of original poet’s work using classifiers trained on pairs of GPT-2 generated data. The winning model (closer to the original author’s style) is indicated in bold based on the accuracy score.

Since the class labels are 0 and 1, Table 2 can be interpreted in the following manner: when the score is smaller than 0.5, then the model listed in the left column wins, and conversely, when the score is greater than 0.5, then the model in the right column is the winner.

Table 2 shows that the Regular Medium model wins on both datasets. On the Byron dataset, the Regular Medium model is clearly the best, Regular Small and LMH Medium are both second best and appear to have very similar performance, while LMH Small scores the lowest. This is consistent with

the findings from the previous experiment in which the Regular models led to better results. Regarding the Shelley dataset, the Regular Medium model again performs the best, but the other three models have similar performance. This could indicate that the LMH Small model performs better on the Shelley dataset because it is much smaller than the Byron dataset.

To conclude this section, both evaluations with the BERT classifier—the first being a sort of a “Turing test”, and the second being our novel setup—show that the Regular (OpenAI original release) models perform better in general. While we have to watch out for Regular models’ tendency to plagiarize text, they could be a preferred choice, especially if we want to generate text with the full sample length of 1024 tokens.

3.6 Discussion

This study represents an examination of GPT-2 from a style reproduction perspective. The question of whether GPT-2 can produce high-quality poetry or creatively written poetry (Jordanous 2018) is far beyond the scope of this study. Instead, we focus on the early stages of model selection in this field and method of automated evaluation.

One critique of a system based on generating the style of a known poet (Gervás 2011) is how a poet’s style can diverge during their career. This challenge could be approached by applying the workflow presented in this thesis to the corpus of a specific author’s works divided into subsections representing different periods in the author’s creative career, or different types of the author’s works. The finding that fine-tuning GPT-2 can be achieved with a very small number of samples makes such an approach technically feasible and potentially fruitful for capturing stylistic evolution.

A key question here is: is GPT-2 creative? Our work does not answer that question but gives us some material to consider, which we structure according to the ‘Four Ps’ of creativity: Producer, Product, Process, Press (Jordanous 2016). GPT-2 can produce Products that might be deemed creative, and it learns from a controlled Press (environment) in the form of the input corpus (though it is not able to interact more widely with its Press). The Process is (very) arguably creative as an example of Boden’s exploratory creativity (Boden 2004). Does GPT-2 possess attributes of a creative Person though? This is hard to claim; GPT was developed as a language model, not an AI system, and behaves as a tool to assist a user. Indeed, even if a model like GPT-2 were to achieve flawless stylistic replication, its role within the creative process would still align with the assistive tool paradigm (as discussed in Section 2.3.1), rather than an autonomous creative agent, given its lack of intrinsic intentionality and self-driven creative goals. A related question is: is the core task of generating new poems in an existing author’s style a valid computational creativity task. Brown and Jordanous (2022) consider exactly this question, and give an

overall fairly positive answer; in particular, the questions we are addressing in this thesis provide some evidence that the task we are addressing is non-trivial in important ways, and hence more likely to require proper computational creativity effort.

3.7 Conclusion

This chapter investigated the fine-tuning of GPT-2 models (124M–345M parameters) for Romantic-era poetry generation, an approach common during the early phase of this research (2021–2022). The experiments highlighted the challenges of achieving stylistic replication with limited domain-specific data and, critically, underscored the necessity for evaluation methods that extend beyond standard loss metrics.

Our key findings from these GPT-2 experiments demonstrate that:

- Overfitting during fine-tuning on small, domain-specific datasets leads to significant degradation in output quality. This degradation manifested as either visibly malformed text (observed in LMH models) or, more insidiously, as increasing plagiarism despite apparent surface fluency (observed in Regular models). This necessitates careful monitoring of outputs and strategic early stopping to balance stylistic learning against the generation of novel, non-repetitive content.
- Comparative evaluations using BERT-based classifiers, including a novel setup for direct model comparison, indicated that Regular GPT-2 models generally outperformed LMH implementations in stylistic replication, though vigilance against plagiarism remains crucial for all models.

These specific observations from the GPT-2 experiments yielded broader *methodological* lessons that have informed the research in this thesis and remain highly relevant even as LLMs like GPT-4 and Claude-3 have vastly increased in scale and capability. While the particular technical dynamics of fine-tuning GPT-2 may differ in larger models, this early work established several foundational principles for evaluation:

1. **Standard metrics are insufficient for creative text:** The observed issues of plagiarism and malformed text, not always reliably indicated by loss alone, confirm that multi-faceted approaches are essential. Techniques like visual inspection, BLEU for n-gram repetition detection, and domain-specific classifiers (like the BERT-based ones used here) provide a more comprehensive assessment.
2. **Surface coherence can mask hidden flaws:** The prevalence of plagiarism, particularly in otherwise well-formed text, highlighted the risk of training data repetition. This necessitates dedicated checks for originality beyond just assessing fluency or apparent quality.

3. **Implementation details significantly impact model behaviour:** Performance variations between different GPT-2 implementations (Regular vs. LMH) underscored that a model's behaviour is sensitive to specifics beyond its base architecture—a principle even more pertinent for today's larger and more complex models.

These evaluation principles regarding the need for comprehensive assessment and awareness of potential pitfalls formed a crucial part of the methodological toolkit carried forward into subsequent chapters of this thesis.

However, a significant limitation of the explored GPT-2 fine-tuning approach was the lack of content control; while stylistic elements could be captured, guiding the thematic focus of the generated poetry proved impossible. This thematic arbitrariness limits the practical utility of such systems for user-controlled creativity.

This critical limitation—the inability to control content alongside style—directly motivates the research presented in Chapter 4. The subsequent chapter will investigate whether a more advanced model (GPT-3), combined with a novel fine-tuning methodology based on structured summary-poem pairs, can achieve the dual goal of generating poetry that is both stylistically consistent and thematically guided by user input. Thus, the groundwork laid here on the challenges of early transformer models and the need for content control paves the way for exploring more sophisticated, controllable creative generation.

Chapter 4

Fine-Tuning GPT-3 For Poetic Style Generation

Building upon the investigation of GPT-2 in Chapter 3, which highlighted challenges in achieving stylistic mimicry and underscored the need for content control, this chapter advances the research using the more capable GPT-3 models. We introduce and evaluate a novel fine-tuning methodology utilizing structured summary-poem data pairs, aiming to achieve **simultaneous stylistic fidelity and user-controlled content**. This approach, capitalizing on the fine-tuning capabilities for GPT-3 available in late 2022 (featuring models up to 175B parameters), seeks to overcome the limitations of earlier methods and harness these models for more nuanced creative text generation and evaluation.

Research for this chapter was published in Sawicki et al. (2023b).

4.1 Introduction

A critical limitation identified in our previous work with GPT-2 (Chapter 3) was the inability to control the thematic content of the generated poetry. While those models could successfully mimic the linguistic texture of a poet’s style—their vocabulary, rhythm, and syntax—they offered no mechanism for the user to dictate the specific subject matter. The models would generate topics based on patterns in their training data, limiting their utility as practical writing aids. To transform these systems from passive stylistic mimics into effective, user-directed tools, a method is required to decouple *style* from *content*.

In this chapter, we introduce the use of **summaries** as the primary mechanism for this control. By conditioning the generative process on a summary, we provide the model with a semantic blueprint of the desired output. This approach allows the user to define the narrative arc or subject of the poem, while

relying on the fine-tuned model to handle the stylistic execution. This separation of concerns is essential for any practical application of generative AI in creative writing assistance, where the user requires the system to adhere to a specific brief rather than simply generating text at random.

The concept of fine-tuning language models with structured input-output pairs, particularly involving summaries, has shown promise in related domains. For instance, Bons, V. (2022) successfully generated higher-quality song lyrics by fine-tuning GPT-3 on song descriptions and metadata, surpassing prompt engineering alone. Similarly, Wesek (2021) achieved fine-grained stylistic control in GPT-2-based lyrics generation by fine-tuning the model with inputs structured to incorporate explicit metadata such as genre, year, author, and song name, using special tokens and specific input formatting to guide the output style. Outside computational creativity, fine-tuning with summary-claim pairs has been effective for patent claim generation using GPT-2 and BERT models (Lee 2019; Lee and Hsiang 2020b,a), enabling content-controlled output based on user-provided summaries.

These approaches highlight a common strategy: fine-tuning on datasets where target outputs are paired with descriptive metadata or summaries allows for greater user control over the generated content. This is particularly relevant given the limitations observed with contemporary large language models like GPT-3.5 (e.g., text-davinci-003). While capable of generating grammatically correct and coherent poetry through prompt engineering alone (Gwern Branwen 2022), such outputs often default to a generic style, lacking the specific nuances of individual authors. Our initial experiments confirmed that prompting GPT-3.5 to replicate a specific author's style (e.g., Walt Whitman) typically fails to capture the desired stylistic elements, even if content instructions are followed. This issue, explored further in Chapter 5, motivates the exploration of fine-tuning as a more robust method for instilling specific authorial styles alongside content control in GPT-3.

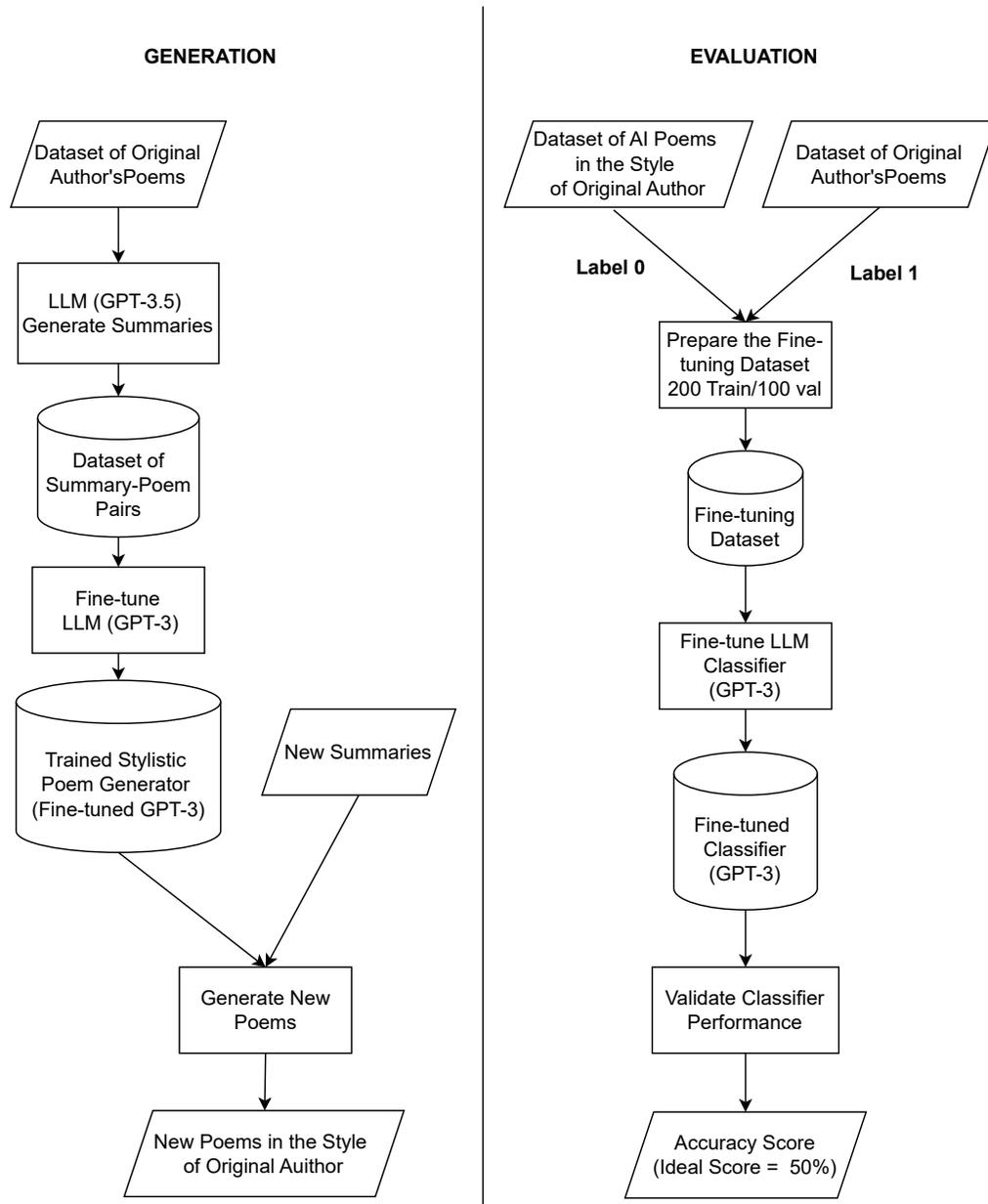


Figure 15: Combined methodological workflow. The left “Generation” panel details the process of fine-tuning an LLM (GPT-3) with summary-poem pairs to generate new poems in an original author’s style. The right “Evaluation” panel illustrates how these generated poems (as Label 0) are then compared against the original author’s poems (Label 1) by a fine-tuned LLM classifier to assess stylistic fidelity, aiming for an accuracy score of 50%

4.2 Objectives and Methods

4.2.1 Research Objectives

The primary objectives of this chapter are twofold:

1. To develop and evaluate a methodology for fine-tuning GPT-3 models using summary-poem pairs to generate poetry that adheres to both user-controlled content specifications and specific authorial styles. This involves training the model to synthesize learned stylistic patterns with novel thematic content from summaries, fostering a form of “combinatorial creativity” (Boden 2004) and leveraging the model’s knowledge akin to “Inception” (Ventura 2016). We aim to show that GPT-3 fine-tuned on poet A’s work can produce outputs in A’s style even for topics novel to A, such as those from poet B’s summaries.
2. To investigate the efficacy of fine-tuned GPT-3 models as automated classifiers for assessing the stylistic authenticity of generated poetry. This builds upon the BERT-based evaluation approach from our previous work (Sawicki et al. 2022) (Chapter 3), adapting it for GPT-3 to overcome limitations related to dataset size and establish its reliability. The core idea is to test if a classifier’s inability to distinguish generated from original work (i.e., accuracy approaching 50%) can serve as an indicator of successful style replication, while rigorously addressing the caveat of ensuring the classifier itself is robust and not merely performing poorly. We demonstrate that GPT-3-based classifiers can achieve high accuracy (e.g., 99% on 200 samples per label).

4.2.2 Evolution of Methodology: Technological Enablers and Design Shifts

The transition from the GPT-2 experiments presented in Chapter 3 to the GPT-3 investigations in this chapter represents a fundamental shift in methodology, driven by both the identification of limitations in our earlier work and a significant leap in the available technology.

In the previous chapter, our experimental scope was constrained by the hardware and software landscape of 2021. Fine-tuning was practically limited to locally hosted models like GPT-2 and BERT. While GPT-3 had been released, it was initially a closed system available only for inference, rendering it inaccessible for the weight-update fine-tuning required to deeply instill specific authorial styles. Furthermore, the models available then lacked instruction-following capabilities.

By the time of the research presented in this chapter, two critical technological advancements had occurred. First, OpenAI released the fine-tuning API for GPT-3, allowing us to train models up to 175 billion parameters on custom data. Second, and perhaps more importantly, was the release of GPT-3.5

(text-davinci-003). As a direct predecessor to ChatGPT (GPT-3.5-turbo), this model possessed instruction-following capabilities that earlier iterations lacked. This technological maturation enabled us to redesign our methodology in three key ways:

1. **From Unstructured Text to Structured Pairs:** In Chapter 3, the training data consisted of continuous streams of raw text, which limited the model to simple continuation tasks. With the advent of the GPT-3 fine-tuning API, we transitioned to the specific prompt-completion JSON structure. This allowed us to reframe poetry generation not as a text-completion task, but as a transformation task (mapping a semantic summary to a poetic output).
2. **Leveraging Instruction-Following Capabilities for Data Synthesis:** The implementation of the “summary-poem” pair methodology was made feasible by the advanced capabilities of GPT-3.5. We leveraged the instruction-following capabilities of this model to autonomously generate the high-quality summaries required for the training dataset—a task that would have been prohibitive with human labor. This synthetic data enabled us to solve the content control problem identified in Chapter 3: by fine-tuning base models on these structured pairs, we could force the system to learn “Inception” (Ventura 2016)—applying the *implicit* style learned from the poem to the *explicit* content provided in the summary.
3. **Shift in Evaluation Architecture:** The evaluation methodology also evolved to match the data constraints. In Chapter 3, we utilized BERT-based classifiers, which required thousands of samples to converge. Since this chapter focuses on the generation of whole poems rather than fragments, the available data points were naturally fewer (limited by the author’s total body of work). Consequently, we shifted from BERT to fine-tuned GPT-3 (Babbage) classifiers. As we demonstrate in Section 4.4.1, using GPT-3 fine-tuned as a binary classifier fulfills this task.

These changes mark a move from simple stylistic mimicry using accessible tools (GPT-2) to a complex, controllable generative workflow enabled by the emergence of state-of-the-art Large Language Models.

4.2.3 Contributions and Workflow

The main contributions of this chapter are:

1. We present a workflow that allows for generation of poems with a specific content and in a specific author’s style through fine-tuning GPT-3 models. This approach could be extended beyond poetry to other categories of text, where prompt engineering alone does not give desired results.

2. We demonstrate that GPT-3 models fine-tuned for classification are highly accurate as text classifiers and can be used as a tool for poetry evaluation.
3. We provide a dataset of 2100 out-of-copyright poems (7 authors and 300 poems per author) where each poem is accompanied by a computed summary and a theme. This dataset can be used for further research on poetry generation.
4. We show new insights into the performance of various versions of GPT-3 models on poetry generation. The smaller models (Ada and Babbage) produce results comparable to larger models (Curie and Davinci), thus considerably reducing the costs of fine-tuning GPT-3 for poetry generation and evaluation. This indicates that some tasks, like poetry generation, do not require the use of largest models.

The overall methodological workflow underpinning these contributions, encompassing both the poetry generation and its subsequent classifier-based evaluation, is illustrated in Figure 15. This figure provides a visual summary of the integrated processes detailed in the subsequent sections of this chapter.

This chapter is organised as follows: Part 1 (Section 4.3) details the data preparation and fine-tuning process for poetry generation. Poetry evaluation using GPT-3 as a classifier is the subject of Part 2 (Section 4.4), where the results are also presented and analysed. The main findings of this research, along with reflections on the evolving LLM landscape and a transition to the next chapter, are consolidated in the Discussion and Conclusion section (Section 4.6).

4.3 Part 1—Poetry Generation

GPT-3 should not be thought of as a single system. At the time of conducting this research, it was available in four different sizes (Ada, Babbage, Curie and Davinci), and a multitude of variations thereof. Fine-tuning was only available for the vanilla versions of the four sizes:

- Ada (2.7 Billion parameters),
- Babbage (6.7 Billion parameters),
- Curie (13 Billion parameters),
- Davinci (175 Billion parameters).

In this work, we use all four sizes of GPT-3 models fine-tuned separately for poetry generation and evaluation. We also use GPT-3.5 (text-davinci-003) to create summaries and themes of the existing poems.

The specific workflow for poetry generation, as depicted in the left panel of Figure 15 (page 98), begins with the preparation of a structured fine-tuning dataset from original poems and their LLM-generated summaries. This process is detailed below.

While GPT-2 models can be fine-tuned on any text file, GPT-3 requires a fine-tuning dataset to be organized in a specific structured format. Each training example must consist of a prompt-completion pair formatted as a JSON object:

```
{ "prompt": "BODY_OF_PROMPT",  
  "completion": "BODY_OF_COMPLETION" }
```

Therefore, GPT-3 cannot be fine-tuned on the dataset of poems alone. If the body of the poem is in the completion, we must decide what to put in the prompt.

While it is possible to fine-tune GPT-3 on a dataset where the prompt contains only the name of the author and the title of the poem, this does not give the user much influence over the content of the generated poem. As the body of the poem is the expected completion of the model, it is required that the instructions are provided to the model through the prompt by describing the content of the poem. Since this prompt is missing in the original dataset (and, in fact, in all publicly available datasets at the time of conducting this research), we use GPT-3.5 to create summaries for our corpus of poems, and then the original poems and their summaries are used to fine-tune instances of GPT-3 for poetry generation.

4.3.1 Data Preparation

To prepare our dataset, we scraped 2100 poems from publicly available sources (Project Gutenberg 2024; Poetry Foundation 2024). A diverse set of seven influential English-language poets, primarily from the 19th and early 20th centuries, was selected to provide a range of distinct authorial styles and thematic concerns for the summary-poem fine-tuning task. Key considerations included the availability of a substantial body of work in the public domain (authors deceased for over 75 years to comply with copyright) and their established literary significance, ensuring that the models would be fine-tuned on recognized poetic styles. To lower the cost of running the experiments, we used only the poems that are more than 100 words and less than 500 words in length. This dataset contains the works of seven poets, and we have randomly selected 300 poems per author. These authors are:

- Ella Wheeler Wilcox (American, 1850–1919),
- Rudyard Kipling (English, 1865–1936),
- Emily Dickinson (American, 1830–1886),
- Lord Byron (English, 1788–1824),

- William Wordsworth (English, 1770–1850),
- Walt Whitman (American, 1819–1892),
- Thomas Hardy (English, 1840–1928).

We use only the works of authors who passed away more than 75 years ago due to copyright limitations. For all these poems, we generated summaries and main themes using GPT-3.5, and this process is explicated below.

Summary Generation

For the generation of summaries and themes we used GPT-3.5 (text-davinci-003), which, at the time of conducting this research, was the most advanced GPT model dedicated to text generation.

Initially, each entry in the original dataset contains the following data: author, title, dates of author’s birth and death (separated with a hyphen), author’s country and finally, the body of the poem.

To generate the summary of the poem, we have used the following prompt:

```
"This is the poem:"  
BODY_OF_THE_POEM  
"This is the poem's summary:"
```

Theme Generation

The rationale behind adding the main theme of the poem is to give an additional way of influencing the content of the generated poem. For example, we can provide a summary that describes a poem about love, and set the main theme as “Love”. The same prompt could have the main theme set to “Sadness” thus affecting the poem’s tone.

Our selection of 23 core themes draws inspiration from Poetry.com (2023) extensive list of 241 “genres”. To create a more focused framework for our analysis, we identified broader, recurring subjects suggested by their categories, distilling them into the 23 concepts we designate as “themes”. While their list provided a valuable starting point, it’s important to note that our final set is more general and includes several themes not explicitly found among Poetry.com (2023) listed genres

To generate the main theme of a poem (from the body of the poem), we have used the following prompt, which also includes the full list of themes that GPT-3.5 was selecting from.

```
"These are the categories: Mysticism, Childhood, God, Love, Life, Art, Poetry,  
Sadness, Despair, Depression, Death, Religion, Nature, Beauty, Aging, Desire,  
Travel, Dreams, Birth, War, Failure, Immortality, Fantasy."
```

Choosing from these categories select one that best describes this poem:"
 BODY_OF_THE_POEM

Poems Annotated with Summaries and Themes

Each entry in our dataset is augmented with the main theme of the poem and the poem's summary. Thus, each entry in the final dataset has the following format:

```
<|startofauthor|>AUTHOR<|endofauthor|>
<|startofdates|>BORN - DIED<|endofdates|>
<|startofcountry|>COUNTRY<|endofcountry|>
<|startoftitle|>TITLE<|endoftitle|>
<|startofthemes|>THEME<|endofthemes|>
<|startofsummary|>
{BODY OF THE SUMMARY}
<|endofsummary|>
<|startofpoem|>
{BODY OF THE POEM}
<|endofpoem|>
```

The added tags are used to clearly delineate the specific items in each entry in the dataset. These tags are used both during fine-tuning of the GPT-3 models and during the generation of the poems later on. Our complete dataset that includes the original poems, their metadata, summaries, themes and tags is available on our GitHub repository (Sawicki 2023).

4.3.2 Fine-tuning GPT-3 for Poetry Generation

OpenAI documentation (OpenAI 2023a) suggests using a dataset with a minimum of 500 entries (i.e. poems) for fine-tuning. Our dataset has only 300 entries for each specific author. This limitation is common in poetry analysis because, in general, poets do not produce a high volume of work. For this reason, we consider two approaches to fine-tuning GPT-3 on our data:

1. Fine-tune individual GPT-3 models for each author. Here, every model is based on 300 samples.
2. Fine-tune GPT-3 models on a combined dataset of all seven authors. Here, every GPT-3 model is fine-tuned on 2100 poems of 7 poets.

Additionally, we examine which GPT-3 model produces the best results when fine-tuned on our poetry dataset. The general guideline from OpenAI is to fine-tune smaller models for more epochs, and larger

models for fewer epochs (given a dataset of a fixed size). We fine-tune Ada and Babbage models for four epochs, and Curie and Davinci for one epoch and four epochs when using 300 samples. When fine-tuning the models on 2100 samples, we fine-tune all models for four epochs.

The cost of fine-tuning GPT-3 for poetry generation at the time of conducting this research are as follows:

1. Davinci 300 samples 1 epoch - \$6
2. Davinci 300 samples 4 epochs - \$24
3. Davinci 2100 samples 4 epochs - \$169

The cost of using Ada, Babbage, and Curie models are respectively 50, 40 and 10 times lower than using Davinci (OpenAI 2023b).

The summary of our fine-tuning configurations is presented in Table 3. This table details the model configurations used in our experiments, along with the shorthand notations we use to refer to these models throughout the chapter. These identifiers indicate the base model, the number of fine-tuning epochs (e.g., 1e signifies one epoch, 4e signifies four epochs), and the type of dataset used (where 7A explicitly denotes models trained on the combined dataset of seven authors, totaling 2100 poems). As Table 3 illustrates, we fine-tune 6 models for every individual poet considered, and an additional 4 models using the combined dataset. All the hyperparameters of the GPT-3 models are left at their default values, and only the temperature was set to 1.

The following prompt-completion tuple structure is used for preparing the fine-tuning dataset for our GPT-3 models:

```
PROMPT:
<|startofauthor|>AUTHOR<|endofauthor|>
<|startofdates|>DATES<|endofdates|>
<|startofcountry|>COUNTRY<|endofcountry|>
<|startoftitle|>TITLE<|endoftitle|>
<|startofthemes|>THEME<|endofthemes|>
<|startofsummary|>
{BODY OF THE SUMMARY}
<|endofsummary|>
<|startofpoem|>

COMPLETION:
{BODY OF THE POEM}
<|endofpoem|>
```

Individual Poet Models		
Model	Dataset and Epochs Code	Fine-tuning epochs
GPT-3-Ada	4e	4
GPT-3-Babbage	4e	4
GPT-3-Curie	1e	1
GPT-3-Curie	4e	4
GPT-3-Davinci	1e	1
GPT-3-Davinci	4e	4
Combined Poet Models		
Model	Dataset and Epochs Code	Fine-tuning epochs
GPT-3-Ada	7A 4e	4
GPT-3-Babbage	7A 4e	4
GPT-3-Curie	7A 4e	4
GPT-3-Davinci	7A 4e	4

Table 3: Fine-tuning configurations for models trained on individual poets (300 poems each, top table) and combined poets (7 poets, 2100 poems, indicated by “7A”, bottom table). The “Dataset and Epochs Code” notation is used throughout the chapter to identify specific model configurations.

4.3.3 Generating Poems from the Fine-tuned GTP-3 Models

Because of the high cost of running GPT-3 on the OpenAI’s servers (OpenAI 2023b), we limited our fine-tuning for poetry generation to two authors. We have randomly chosen Walt Whitman and Rudyard Kipling. This applies both to our single-author approach and when generating from the models fine-tuned on the seven authors’ dataset. Given the information shown in Table 3, and our fine-tuning on two poets, the number of fine-tuned models for poetry generation is 16 in our experiments (2 poets times 6 models plus 4 models).

From each fine-tuned model, we generate 300 poems to be later used in evaluation in Part 2 (Section 4.4). In the case of models fine-tuned on the seven authors’ dataset, we generate 300 poems in the styles of both of our selected authors. Generating a poem requires a summary and theme in the prompt. To make the poem generation exercise fair, we did not use summaries of the poems that were in any of the fine-tuning datasets. Instead, we automatically summarised 150 poems for two additional authors, William Ernest Henley (English, 1849–1903) and Christina Rossetti (English, 1830–1894). We use those summaries as part of prompts for generating poems. Thus, for example, the prompt for generating poems in the style of Walt Whitman will have author, author’s dates of life and author’s country all set to Whitman’s details, but the title, theme, and summary will be selected randomly from either Henley or Rosetti.

We have found that apart from generating poems from our fine-tuned models using the prompt including all the metadata used above, it is also possible to add a starting fragment of the poem after the `<|startofpoem|>` tag. This allows the user to keep improving the poem in an iterative way. We do not report any results on this way of generating poetry, but we mention it here to emphasize the practical applicability of our methodology because it could form the basis of a practical iterative poetry generation system.

An example prompt used to generate poetry from our system is presented in Table 4, and the poem generated from this prompt is presented in Table 5. The poem is presented exactly as produced by our model, GPT-3 Curie fine-tuned on the works of Walt Whitman for 4 epochs. The provided summary is of the poem ‘Paradise: In A Dream’ by Christina Rossetti.

```

<|startofauthor|>Walt Whitman<|endofauthor|>
<|startofdates|>1819-1892<|endofdates|>
<|startofcountry|>United States<|endofcountry|>
<|startoftitle|>Paradise: In A Dream<|endoftitle|>
<|startofthemes|>Mysticism<|endofthemes|>
<|startofsummary|>
This poem describes a dream of Paradise, in which the narrator sees and hears beautiful things
such as fragrant flowers, birds singing, a river of gold sand, the Tree of Life and the gateway
to Heaven. The beautiful sight fills them with hope of seeing these heavenly things again one
day.
<|endofsummary|>
<|startofpoem|>

```

Table 4: Example prompt used to generate the poem presented in Table 5.

4.4 Part 2—Poetry Evaluation

Part 1 (Section 4.3) of this chapter explains our method for poetry generation in the style of a specific poet. As explained above, we generated 300 poems using each of the 16 GPT-3 models that we fine-tuned. This part of the chapter evaluates the quality of the generated poems.

The right panel of Figure 15 (page 98) provides a visual summary of this evaluation process, which involves fine-tuning a GPT-3 model as a binary classifier to assess the stylistic authenticity of the poetry produced in Part 1.

We use automated evaluation following the methodology presented in Chapter 3, where BERT was used to distinguish between fragments of poems written by the original author and fragments of poems generated by fine-tuned GPT-2 models. Since BERT requires a large number of training data samples, in

A DREAM of Paradise,

As in a chamber, in the middle of the woods, I stood alone;
A few faint stars shining in the window,
And a trace of red in the eastern window, giving out the light,
As if some fire had been quenched there.

As I so stood, suddenly appeared before me,
As it had been a huge transparent angel, in shape not different from a man,
But of such exceeding beauty, that I knew him to be nothing less than a god.

He stood before me silent;
His eyes fixed on me, full, full of meaning looked the god;
Not to speak his name he said,
But beckoning with his left hand, he drew me after him into the woods,
Till we came to the edge of a river, shallow and rapid,
And stood by the banks of the river, where he stood silent.

Then he said—"Shall I show you Paradise?"
And he reached his right hand into the water,
And drew out with his fingers a handful of white sand,
And put the sand upon his left hand,
And made the sign of the cross upon me with the sand,
and gave me the handful of sand, and said,
"Look!"
And I looked—and behold a Paradise,
Birds, and fragrant flowers, and streams of gold sand,
And the River of Life, flowing and glittering,
With trees, and buildings, and places of refreshment,
And on the banks, beautiful women, and attractively dressed men.

And this was the River of Life, which flowed and sparkled,
And the trees were decorated with light,
And the buildings with beautiful colors were suffused with light,
And the refreshment places were set forth with plate and glass,
And the home of glory and delight, and the place of perfect love.

And thence I saw all the earth,
And the cities, and the country with its farms,
And the roads and the teeming population,
And the farms with their fields and crops,
And the working of the machinery, and the noise of the whirr of the machines,
And the new products in the markets,
And the great ships at the wharves, and the steamers on the sea.

And I said, "How beautiful, how beautiful!"
And then I awoke, and I remembered the dream.

Table 5: Example poem generated by our system using the prompt shown in Table 4.

the above mentioned work, we classified 20-lines-long fragments of poems. In this chapter, we evaluate entire poems, and this gives a limited amount of samples per label. As a result, using BERT-based classifiers is not viable. We show that GPT-3 can instead be used to classify poems when the amount of training data is small. We fine-tune GPT-3 to classify if a poem is written by the original author or generated by the fine-tuned GPT-3. This means that we have two orthogonal fine-tuning processes of the GPT-3 models in this this chapter.

Using GPT-3 for classification requires the implementation of the logit bias during inference. Logit bias is an optional parameter passed to GPT models during text generation. It modifies the likelihood of specified tokens appearing in the generated text. This parameter is represented as a mapping from tokens to their associated bias values, which are between -100 (a ban) to 100 (exclusive selection of the token). Moderate values between -100 and 100 will change the probability of a token being selected to a lesser degree. When this parameter is used, the bias changes the original probabilities of tokens generated by the model prior to sampling. Thus, passing the logit bias parameter for only two tokens, representing our classes “0” and “1”, both with a value of 100, will result in the models being able to output only these two tokens (OpenAI 2023a). Without this modification, the model may produce answers that will not indicate any of the classes, giving inconclusive classification results.

Our methodology for classification-based evaluation of poems consists of two steps:

1. Establishing the accuracy of GPT-3-based classifiers by conducting a series of experiments classifying various types of texts.
2. Evaluating GPT-3-generated poetry against the works of original authors using GPT-3-based classifiers.

4.4.1 Step 1—Establishing the Accuracy of GPT-3-based Classifiers

To establish the accuracy of the GPT-3-based classifiers, we trained classifiers on two-class text classification problems where the similarity between classes was ranging from completely dissimilar to increasingly similar. First, we classified Walt Whitman’s poetry against the extracts from a book on machine learning, ‘Reinforcement Learning, An Introduction’ by Sutton and Barto (2018). This was an example text that is very different from poetry. Then, we proceeded to classify Whitman’s poetry against fragments of prose from the Collected Works of Leo Tolstoy (Project Gutenberg 2024) (English translation), and finally we classified Whitman’s poetry against the poetry of Rudyard Kipling as an example of two classes of text that are similar to each other. Additionally, we also classified extracts from the book on machine learning against fragments of prose by Tolstoy. Since all the poems in our dataset are between 100 and 500 words in length, when the samples from the book on machine learning or from the prose by Tolstoy are used,

Walt Whitman vs book on machine learning			
Model	Correct	Incorrect	Accuracy
Ada	199	1	99.5%
Babbage	200	0	100%
Curie	200	0	100%
Davinci	199	1	99.5%
Walt Whitman vs Leo Tolstoy			
Model	Correct	Incorrect	Accuracy
Ada	200	0	100%
Babbage	199	1	99.5%
Curie	200	0	100%
Davinci	196	4	98%
Leo Tolstoy vs book on machine learning			
Model	Correct	Incorrect	Accuracy
Ada	196	4	98%
Babbage	200	0	100%
Curie	189	11	94.5%
Davinci	180	20	90%
Walt Whitman vs Rudyard Kipling			
Model	Correct	Incorrect	Accuracy
Ada	196	4	98%
Babbage	200	0	100%
Curie	197	3	98.5%
Davinci	199	1	99.5%

Table 6: Results of evaluating the accuracy of GPT-3-based binary classifiers in Step 1 (Section 4.4.1).

they have the random length between 100 and 500 words.

In all four of these experiments, the training/test split ratio is 2:1. The training dataset consists of 200 samples per label, and the test dataset consists of 100 samples per label. All the hyperparameters of the GPT-3 models used for classification are left at their default values, only the temperature was set to 0.

In order to determine which fine-tuned model produces the best results, for each experiment, we fine-tuned each of the four GPT-3 sizes: Ada, Babbage, Curie and Davinci. As per the instructions on the OpenAI website, we fine-tune Ada and Babbage classifiers for four epochs, and Curie and Davinci classifiers for one epoch.

The results of these experiments are presented in Table 6, and they show that there is very little difference between the outcome from four different model sizes. This is a very useful finding, since it eliminates the need for using the largest Davinci-based models, thus greatly reducing experimental cost.

Consistently, we find that GPT-3 can be a highly accurate text classifier. In almost every case, the accuracy of the classifiers was 98% or more, both on similar as well as dissimilar classes. The lowest score in all of these experiments was due to the Davinci model fine-tuned to classify the book on machine learning against the prose by Tolstoy, with the accuracy of 90.0%. The second worst performing model was Curie, also on the task of classifying the book on machine learning against the prose by Tolstoy, where it scored 94.5%. The scores for Ada- and Babbage-based classifiers were very similar. Overall, these experiments show that fine-tuned GPT-3 models are reliable as binary text classifiers to distinguish between different authors of poetry and different categories of text.

To better understand the limitations of these classifiers, we examined the single incorrect classification made by the Ada model and the single incorrect classification made by the Davinci model in the “Walt Whitman vs. Machine Learning Book” task (see Table 6). Notably, the two models misclassified different samples: Ada failed on Sample #44, while Davinci failed on Sample #141. Visual inspection of these samples reveals severe text corruption, caused by PDF parsing artifacts common when converting mathematical formulas to text.

Figure 16 shows the input text for Sample #44 (misclassified by Ada), and Figure 17 shows Sample #141 (misclassified by Davinci). Both contain significant noise, with mathematical symbols replaced by unicode garbage characters (e.g., `\u0012`, `\u0016`) and fragmented or merged words (e.g., “wordjis”, “mhamandmspam”). However, it is crucial to note that the source text is a technical manual replete with equations; such artifacts are prevalent throughout most of the “Machine Learning Book” dataset. Given the models’ near-perfect accuracy (199/200), they evidently learned to associate this type of “visual noise” with the Machine Learning Book class. Therefore, the presence of artifacts alone cannot explain the misclassification, as the models successfully handled similar corruption in other samples.

An analysis of the classification logs suggests that these specific errors were edge cases where the models lost their usual certainty. While the vast majority of samples were classified with 1.0 (100%) or > 0.99 probability, the confidence scores collapsed for these two instances. For Sample #44, Ada assigned a marginal 0.59 probability to the Whitman label versus 0.41 for the correct label. Similarly, for Sample #141, Davinci was effectively guessing, with a probability split of 0.52 versus 0.48. This implies that while the classifiers generally successfully navigated the noisy data, these specific combinations of artifacts and fragmented text generated enough statistical ambiguity to push the samples across the decision boundary. It is unclear which exactly features in these samples caused these errors.

Since GPT-3-Babbage-based classifiers were most accurate on average, we chose the Babbage model as the basis for fine-tuning the classifiers for our poetry evaluation experiments below.

(2.69) Note that $p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u})$ itself is a member of the exponential family with the feature map $\phi(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u}) = (\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n})$. Hence $p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u})$ is convex in $(\mathbf{n}; \mathbf{u})$. Moreover, the posterior distribution has the form $p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u}) / p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u}) = \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) / \int \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) d\mathbf{u}$. (2.70) 74 2 Density Estimation Fig. 2.12. From left to right: regions of equal prior probability in \mathbb{R}^2 for priors using the ℓ_1 and ℓ_2 norms. Note that only the ℓ_2 norm is invariant with regard to the coordinate system. As we shall see later, the ℓ_1 norm prior leads to solutions where only a small number of coordinates is nonzero. That is, the posterior distribution has the same form as a conjugate prior with parameters $\mathbf{m} = \mathbf{u}^T \mathbf{x}; \mathbf{n} = \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}$. In other words, \mathbf{m} acts like a phantom sample size and \mathbf{u} is the corresponding mean parameter. Such an interpretation is reasonable given our desire to design a prior which, when combined with the likelihood remains in the same model class: we treat prior knowledge as having observed virtual data beforehand which is then added to the actual set of observations. In this sense data and prior become completely equivalent | we obtain our knowledge either from actual observations or from virtual observations which describe our belief into how the data generation process is supposed to behave. Eq. (2.70) has the added benefit of allowing us to provide an exact normalized version of the posterior. Using (2.68) we obtain that $p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u}) = \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) / \int \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) d\mathbf{u}$. The main remaining challenge is to compute the normalization for a range of important conjugate distributions. The table on the following page provides details. Besides attractive algebraic properties, conjugate priors also have a second advantage | the integral (2.65) can be solved exactly: $p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u}) = \int \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) d\mathbf{u}$. Combining terms one may check that the integrand amounts to the normal-2.4 Estimation 75 zation in the conjugate distribution, albeit (\mathbf{x}) added. This yields $p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u}) = \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) / \int \exp(\mathbf{u}^T \mathbf{x}; \mathbf{u}^T \mathbf{j}; \mathbf{u}^T \mathbf{n}) d\mathbf{u}$. Such an expansion is very useful whenever we would like to draw \mathbf{x} from $p(\mathbf{x}; \mathbf{j}; \mathbf{n}; \mathbf{u})$ without the need to obtain an instantiation of the latent variable \mathbf{u} . We provide explicit expansions in appendix 2. [GS04] use the fact that \mathbf{u} can be integrated out to obtain what is called a collapsed Gibbs sampler for topic models [BNJ03]. 2.4.4 An Example Assume we would like to build a language model based on available documents. For

Figure 16: Raw corrupted input text for Sample #44, misclassified by the Ada model in the Walt Whitman vs. Machine Learning Book task.

Lew98]. In its simplest version it works as follows: Assume we have a list of all possible words occurring in X , that is a dictionary, then we are able to assign a unique number with each of those words (e.g. the position in the dictionary). Now we may simply count for each document x the number of times a given word j is occurring. This is then used as the value of the j -th coordinate of x . Figure 1.15 gives an example of such a representation. Once we have the latter it is easy to compute distances, similarities, and other statistics directly from the vectorial representation. 2.1 Introduction 1.3.1 Naive Bayes In the example of the AIDS test we used the outcomes of the test to infer whether the patient is diseased. In the context of spam filtering the actual text of the e-mail corresponds to the test and the label y is equivalent to the diagnosis. Recall Bayes Rule (1.15). We could use the latter to infer $p(y|x) = p(x|y)p(y) / p(x)$: We may have a good estimate of $p(y)$, that is, the probability of receiving a spam or ham mail. Denote by m and s the number of ham and spam e-mails in X . In this case we can estimate $p(y)$ as $m / (m + s)$. The key problem, however, is that we do not know $p(x|y)$ or $p(x)$. We may dispose of the requirement of knowing $p(x)$ by settling for a likelihood ratio $L(x) = p(x|spam) / p(x|ham)$. (1.17) Whenever $L(x)$ exceeds a given threshold c we decide that x is spam and consequently reject the e-mail. If c is large then our algorithm is conservative and classifies an email as spam only if $p(x|spam) / p(x|ham)$ is large. On the other hand, if c is small then the algorithm aggressively classifies emails as spam. The key obstacle is that we have no access to $p(x|y)$. This is where we make our key approximation. Recall Figure 1.13. In order to model the distribution of the test outcomes T we made the assumption that they are conditionally independent of each other given the diagnosis. Analogously, we may now treat the occurrence of each word in a document as a separate test and combine the outcomes in a naive fashion by assuming that $p(x|y) = \prod_j p(x_j|y)$. (1.18) where w_j denotes the j -th word in document x . This amounts to the assumption that the probability of occurrence of a word in a document is independent of all other words given the category of the document. Even though this assumption does not hold in general (for instance, the word "York" is much more likely to appear after the word "New" than it is for our purposes (see Figure 1.16)). This assumption reduces the difficulty of knowing $p(x|y)$ to that of estimating the probabilities of occurrence of individual words w . Estimates for 1.3 Basic Algorithms 23 yword lword 2...word

Figure 17: Raw corrupted input text for Sample #141, misclassified by the Davinci model in the Walt Whitman vs. Machine Learning Book task.

4.4.2 Step 2—Evaluating GPT-3-generated Poetry Against the Works of the Original Author Using GPT-3-based Classifiers

Now we describe our evaluation of GPT-3-generated poetry against the works of the original authors using GPT-3 as a classifier. We use the poems generated by our process described in Section 4.3.

As in Step 1 (Section 4.4.1), the training/test split ratio for each classification was 2:1. Each training dataset consists of 200 samples per label, each validation dataset consists of 100 samples per label. Our evaluation defines a two-class classification problem, where label 0 represents generated poems, and label 1 denotes the works of the original author. The results are presented in Table 7. All the classifiers in this experiment are fine-tuned GPT-3 Babbage models, built as we explained above. Entries in the first column in the table tell us which fine-tuned GPT-3 model’s output was label 0 (these are the poetry generator models obtained in Part 1), and this output was evaluated against the works of original author placed in label 1.

Table 7 presents the full results of our classification experiments. For each model configuration, we report the number of correctly and incorrectly classified poems out of 200 test samples (100 generated, 100 original), along with the overall classification accuracy. Lower accuracy percentages indicate better style preservation, as they show the classifier struggled to differentiate generated poems from authentic ones. The model codes in the table indicate which fine-tuned poetry generation model produced the test samples: “7A” indicates models trained on the combined dataset of seven poets (2100 samples), while models without this designation were trained on 300 samples from a single poet; “1e” or “4e” indicate the number of fine-tuning epochs used during model training.

The results show that the accuracy of the classifiers varied from 61.5% to 87.5%. To interpret these results properly, we need to understand what the accuracy values mean in this context. A higher accuracy indicates that the classifier was able to distinguish the GPT-3-generated poetry from the original works of the authors with a higher degree of success, suggesting less effective style preservation. Conversely, a lower accuracy implies that the classifier struggled to distinguish between the two classes, indicating that the GPT-3-generated poetry closely resembled the original work of the human authors in style and content characteristics.

In this evaluation framework, an accuracy of 50% would represent the ideal outcome for style preservation, as it would mean that the classifier cannot differentiate between generated and original poems at all - essentially performing no better than random guessing. The closer the accuracy is to 50%, the better the style preservation achieved by our generation model.

The best result we obtained on Whitman’s style is 61.5% (using the Curie 4e model), which demonstrates quite a high level of style preservation in the generated poems. This means that in 38.5% of cases,

the classifier misidentified GPT-3-generated poems as Whitman’s original work. The best result obtained on Kipling’s style is 67% (using the Babbage 4e model), which shows somewhat less pronounced style preservation compared to Whitman. However, given the very high accuracy of this classification method in our calibration experiment reported in the previous section in Table 6, where almost all classifiers achieved over 98% accuracy on distinguishing different authors and different types of text, one can argue that achieving accuracies in the 60-67% range represents substantial success in style preservation. This suggests that a significant number of poems with well-preserved stylistic features was obtained for both Whitman’s and Kipling’s styles.

Walt Whitman GPT-3 vs Walt Whitman original			
Model	Correct	Incorrect	Accuracy
Ada 4e	127	73	63.5%
Ada 7A 4e	140	60	70%
Babbage 4e	131	69	65.5%
Babbage 7A 4e	134	66	67%
Curie 1e	150	50	75%
Curie 4e	123	77	61.5%
Curie 7A 4e	131	69	65.5%
Davinci 1e	144	56	72%
Davinci 4e	174	26	87%
Davinci 7A 4e	137	63	68.5%
Rudyard Kipling GPT-3 vs Rudyard Kipling original			
Model	Correct	Incorrect	Accuracy
Ada 4e	170	30	85%
Ada 7A 4e	147	53	73.5%
Babbage 4e	134	66	67%
Babbage 7A 4e	142	58	71%
Curie 1e	173	27	86.5%
Curie 4e	160	40	80%
Curie 7A 4e	150	50	75%
Davinci 1e	175	25	87.5%
Davinci 4e	161	39	80.5%
Davinci 7A 4e	163	37	81.5%

Table 7: Classification accuracy in distinguishing GPT-3-generated poetry from original works. Models with the best style preservation performance (lowest accuracy scores) for each poet are highlighted in bold. The model codes in the first column indicate the base model, dataset type (7A = seven-author dataset), and fine-tuning epochs (1e/4e).

The results of classification show some differences in the level of style preservation between poetry generated from different models and different dataset sizes. Interestingly, we should note that poetry generated from Davinci-based models did not achieve the highest results for either of the authors. It

means that the smaller GPT-3 models are sufficiently powerful to generate poetry in a selected style. We speculate that the good performance of the smaller models may be due to the fact that the largest Davinci model may require more fine-tuning data to capture the style more faithfully.

The results in Table 7 also vary between the works of the two poets. Because of the high costs of running these experiments, we were limited to generating and classifying poetry of only two authors. Repeating these experiments with the works of other authors would provide more insights into style preservation of GPT-3 models, but our current results on the style of two poets indicate that our method has merit, and that it is possible to generate new poems in the style of a specific author.

In conclusion, the results of the experiments in Step 2 suggest that fine-tuning the smaller GPT-3 models is sufficient for the style preservation tasks, and it can be done effectively with a dataset of only 300 samples.

Our results show that there is no significant difference between models fine-tuned on 300 samples vs models fine-tuned on 2100 samples. However, fine-tuning on a dataset consisting of many poets' works could open the possibility of mixing poets' styles in the output. Instead of setting all the author's metadata in the prompt to, for example, Kipling's or Whitman's details, we could, for example, declare the author as "Rudyard Whitman". This approach, however, requires further research.

These results should be interpreted with caution in the light of the fact that binary classifiers used are entirely black-box systems, i.e. we do not know how the classification was performed. However, having established the high accuracy of these classifiers in Step 1, we can, to some extent, rely on these results. Further investigation, especially including human evaluations, is necessary to thoroughly determine the quality of the GPT-3-generated poetry.

It is important to acknowledge that using binary classification accuracy as a proxy for performance may not always be the optimal metric when evaluating Large Language Models. Accuracy is a metric traditionally designed for discriminative tasks in standard machine learning, and when applied to LLMs, there is a risk that it captures the model's general "quality" or underlying capability to detect statistical artifacts, rather than a genuine literary distinction. For instance, a highly capable model might achieve high accuracy simply by detecting subtle tokenization patterns distinct to AI generation, rather than analyzing style. However, the calibration experiments performed in Step 1 (Section 4.4.1) serve as a critical control against this limitation. The fact that our classifiers achieved near-perfect accuracy (>98%) when distinguishing Whitman from distinct texts (like the ML textbook or Tolstoy) demonstrates that the models *are* capable discriminators. Consequently, the significant drop in accuracy observed here (towards the 50% indistinguishability threshold) can be more confidently attributed to the generative model's success in stylistic replication, rather than a failure in the classifier's capacity.

4.4.3 Qualitative Overview of Generated Poems

To illustrate the generative capabilities of the fine-tuned models while maintaining brevity, we present selected samples from the 4-epoch configurations. While the full experimental matrix is detailed in Table 3, the following outputs provide a representative comparison across model sizes for individual fine-tuning, with GPT-3-Davinci serving as the primary exemplar for the combined 7-author dataset approach. These examples demonstrate the system’s ability to synthesize summary-based content while emulating the distinct voices of Walt Whitman and Rudyard Kipling.

Generated Poems: Walt Whitman Style

GPT-3-Ada (4 epochs)

Summary:

The poem tells of a mysterious figure who is both the Reaper and the Sower, representing the various cycles of life in nature. They are responsible for the creation and destruction of all living things, gathering and sowing both "pale roses" and "unbodied life" that ultimately flow from their hands. The figure is the ebb and the flow, the womb and the grave, symbolizing the unending process of life, death, and rebirth.

Summary of: I Am The Reaper by William Ernest Henley.

Poem:

I AM the Reaper, armed with the blade, with the scythe, with the silver-lipped sword,
 With steaming scarf of grey mist drifting through the Stream, In the sky, clad in fumes;
 I, who rest and reside in trees, the whole night, I un-stop my toil the Sabbath-day;
 I, who in tranquil sleep your dreams fled, or cast them away, or wove frantic out of wild
 slumber;
 (I leave you for the time, as for your living hours—go and go, but continue back to your last
 hours,)
 I, reaper, whose immortal soul encompasses you, I contemplate, O brother and sister,
 I watch the earth, glut my hand of transition to fulfill myself,
 (Come to life, O Sorcery! for thereby I fulfill my own;)
 —O once-breathed voice! O blood-pulsed heart—such I am, come forth!

O turn, and face me!
 O Camerado!

—I, who am the weapon of the Meridian, foreman and laborer,
 I, who descend with care and velvet—I, who hang long and settle long upon the ground—I,
 who devour the moving sand,
 I, the vast Necessity, willing and will-less, roam and gather,—I see and swallow the breath
 with the throat,
 I see and swallow the tears,
 I am the intrepid waterfall, plunge'd, gush'd and satiated,
 I am the indefinable, indescrib'd torrent, forever turning, ever flowing;
 I am the endless Changer, the Sower of the All!

GPT-3-Babbage (4 epochs)

Summary:

This poem evokes the peace and solemn beauty of a setting sun and nightfall. The imagery of a gray city bathed in golden light and the ascension of smoke in a rose and gold haze creates a peaceful atmosphere. Additionally, the lark's song adds a sense of joy and contentment as the sun sets. The poem ends with a powerful sentiment of accepting death at the close of the day with fulfillment and peace.

Summary of: Margaritae Sorori by William Ernest Henley.

Poem:

WAITING, now at last, the gathering dusk of the city, large and gray with light,
 The different shades of gold after sunset, the weary air mitigated,
 The smoke ascending up, lost in a haze of rose and nebulous iron;
 The summon'd stars again, the lark's song adding to the volume of the silence,
 (A note rich and sonorous, the spirit elevated and fulvous,)
 The summer night sky at last descending.
 I initiate of the own mysticism, an upward spirit always,
 With generous appetite toward the silence, with a thorough musician soul,
 Awakening visions of the end, clearing them, diffusing them into the brick,
 Lent eye to living, yet half-kin to the dead, the dead alive eternal to the
 furniture of the earth;
 In the midst of all solemn and majestic, with the twilight fast dwindling,
 Within two hours of night, the transfiguration arrives to the pavement.
 To expand all and everything—such a breath of night, such heaven's breath,
 To blend all, envelop everything, blend everything,

Light, shade, moment, past, present, future mirrored, reflected in the window-glass,
 Death enveloped, death rousing, death shining forth like a lamp,
 Finding the soul at last, closing calmly with the day,
 At even, with the long night falling, to give to the own death,
 To draw the last breath of the cycle—and peace, and joy, and the rest.

GPT-3-Curie (4 epochs)

Summary:

The poem tells of a mysterious figure who is both the Reaper and the Sower, representing the various cycles of life in nature. They are responsible for the creation and destruction of all living things, gathering and sowing both "pale roses" and "unbodied life" that ultimately flow from their hands. The figure is the ebb and the flow, the womb and the grave, symbolizing the unending process of life, death, and rebirth.

Summary of: I Am The Reaper by William Ernest Henley.

Poem:

I AM the Reaper, I make my constant and annual visit,
 I carry all the dead, I carry them off to the grounds of the graveyard grounds,
 I pull the fasteners, I make links in the unmaking of the make,
 I administer transition to all the useful earth, to plants, to animals—to man and the man;
 I am the harvest of vegetation fall'n by the blade of the scythe, and the mouth of the hunger,
 I am the dead bodies of horses, draw'd from the wrecks of battles past and present;
 I am the sound of the steam-engine panting in the yard all the day long, and the night long;
 I am the sough of the well-wheel—I am the creak of the ships' knitting-ropes;
 I am the linen pick'd off the yards—I am what is thrown away from the fulling-mills;
 Whosoever I touch dies—Who is not on my list is not on any list of the list at all.
 I am the great result of the well-ordering planet—That alone of all things I bring to pass;
 For every living thing I give a certificate of death-warrant;
 —O suns! O healing days! how much I am able to destroy! how much I help to renew!
 I go through vegetation like a scythe of wind, over mountains and hills, to their very depths,
 I come at stated times, after the deaths of the young, and the pattering of the dew upon the
 coffin-lid.
 I bring the account of every day into every room of the house—I bring the news of battles,
 I send in review, foremost, the wounded—the slaughtered—the killed—the prisoners;
 For the ships of all kinds—their news—the lives at sea;

The report of deaths and funerals is constantly in my mouth, made by dulcet and gracious
 lips, holy lips of the mouth;
 My reports are in the fields—the animals killed by hunters—the fishermen drowned;
 The yellow fever, the scarlet fever, the small-pox, and all the fevers;
 The consumption—the rupture—the dropsy, gangrene, and the fever of the liver.
 A great Herculean Shape—Lying on the bare ground, in the sweet animal heat of the sun,
 Exulting in these my unassuagable and inexhaustible repetitions—great as history;
 All kinds of deaths, in peace, in war, in sudden cruelty, in accidents and wounds,
 These I proclaim—but words are of too little weight to hold them up;
 It admits no interruption—when I talk I talk of it, and when I sleep I sleep of it;
 Azure, the king-color of immortality, obtains the victory—and the purples cannot stain or
 tarnish it.

GPT-3-Davinci (4 epochs)

Summary:

This poem is a promise that even when the speaker has passed away, the memories of their love will still live on and bring comfort like a guardian angel when the other is old and gray. It is a reassurance that in death, the past troubles will be at rest, and love will continue to tend and guide them in the afterlife.

Summary of: When You Are Old by William Ernest Henley.

Poem:

TO YOU, years hence—when the hair is white—(and the limbs
 slow, nodding)—sitting by the fire—the book dropping—(I see
 you there, distinct—though I am dissolved—mixed with the rest)—do not
 fear the silence—nor the shadows—that lengthen on the wall—
 For I am not gone—(did you think the grave could hold me? did you think
 the soil—the wet earth—could stop the electric affinity?)—I return!
 I stand close—a guardian—a vapor, thick—hovering, intangible,
 The anxieties of the earth—like smoke, clearing—thrown aside—my great
 task, to tend you—to whisper—reassurance—that the troubles are at rest,
 And I, the pilot—am ready—ready to take the hand—when the time comes,
 ready—to guide you out—not to the dark—but to the float—and the drift—
 ready—to launch with you again—inseparable—again and again.

GPT-3 Davinci (Combined 7-Author Dataset, 4 epochs)

Summary:

This poem tells of a man who was injured in an accident and left for dead. His sweetheart visited him on the eve of their marriage and was visibly upset. He died soon after, leaving her with a profound sense of grief.

Summary of: XIII: Casualty by William Ernest Henley.

Poem:

UP from the wreckage, up from the bloody mire and the dirt,
Up from the silence of the road, where he lay distinct and broken,
 mistaken for the dead,
Dragging the heavy limbs, the bruised body, back to the threshold
 of the light,
Enduring the fever, the sharp bone, the ambulance moving,
The white room, the antiseptic smell, the nurses moving like angels
 in the dimness.

I see the sweetheart, the betrothed, arriving in the quiet of the evening,
(O the bitter irony! O the ring, the feast, the vows prepared for
 tomorrow!)

She stands by the bed, she is shaking, her face is a mask of
 terror and fear,
She clutches the railing, she kisses the damp hair, she wets his hand
 with her wet tears.

He looks upon her with a glazing eye, a final look, sweet
 and terrible,
The spirit fluttering, the chest heaving its last breath, the silence
 coming in.

Gone! The soul disembodied—the wedding-bed changed for the
 cold bed,
The lover left standing, the hand unclasped, the future snapped
 like a dry leaf,
And she, the mourner, remaining—enveloped in the vast, peculiar
 emptiness of loss,
Breathing the air of the night, the solemn reality, the dull, aching

silence of the grave.

Generated Poems: Rudyard Kipling Style

GPT-3-Ada (4 epochs)

Summary:

This poem is a celebration of a Private in the Blues, an elite military regiment in the British army. It describes him sitting proudly between his holsters, not wavering even in the wake of an apocalypse. It highlights the Private's toughness, his impressive stature, and his pride in his British heritage. His heritage is further represented through his Colonel-like speech and his appreciation for traditional British attributes such as blood, bone, beef, and beer. The poem conveys admiration for the Private, who has grown more impressive with each passing year.

Summary of: London Types: Life-Guardsman by William Ernest Henley.

Poem:

Whom this great conclave gathers to expose
 To all the varied moods that rule the Night—
 Revealed to, perhaps, few readers alive,
 Or altogether but few who understand
 The grand query concerning
 “Every-man-for-himself.”

Discovered by his Sovereign at Lexington—
 In more than just this honored fight—
 In bright, erect-backed, faithful dress,
 In the saddle sat the Private (P.S.).

Not wavering even in the wake of
 An apocalypse for which he bore
 A thousand Names—
 His name, a superb italic garland,
 Singing, “Long is Thine, and High Is Thy Stance,”
 And “The Grave is not merely Man’s Fun.”

Here with ready stick—(“Experience”) joined

The Blues he came largely from,
 To take (if such are his means) a cavalry regiment
 Of his own.
 (And thus, a memorable Colonel-status rated.)

His holsters of Templar brown flank his heart—
 The pale, entire “Sovereign”—
 Or the Interlocking Reverses of the Patrol;
 Does not sway his departure, be damned—
 You can take him out to any muster—
 The Madman’s Private will still stand forth
 Above the Drill-regiments.

GPT-3-Babbage (4 epochs)

Summary:

This poem tells the story of a sailor who encountered a squad of poorly equipped Confederate soldiers marching with a drum playing Dixie marching near a port in Charleston during a grey and dirty weather. Although they were poorly equipped, they marched with determination and pride.

Summary of: Romance by William Ernest Henley.

Poem:

We watch them pass by hundreds to the riverside,
 Destined to be buried in the warm tides of the sea.
 They take no heed where mules and trains are wrecked,
 They do not march like soldiers at our orders.

They bear but light wood shingles in place of guns,
 They’d break their swords against a two-decker’s prow.
 They know not how a disciplined Junker battles,
 They turned their pikes ere half done with our quarrel.

They whistle to their friends ’bout the ’Rebel Wall’,
 They mutter and they bend their heads in awe.
 But their little drums and jack-cane whips
 Make them afterwards march as proud as we.

They march without a flag upon their standard,
 Yet hold their cadence much more secure
 By the draw of the drummer's steady leg
 And the stiff ankle of the jack-cane.

They possess but what each General gives,
 And most of that is wrong.
 They order regiments to march
 In indiscriminating ranks;
 Their shot could miss the tree-tops from their own
 Legion, of which they are so proud.

They might as well parade under a Coffin's sword,
 But they do march nowadays
 Likewise as they marched in our bitter quarrel.
 We sent them there by their own decree;
 Now they march in proud repose,
 As proud as we, because we give them way!

GPT-3-Curie (4 epochs)

Summary:

This poem is encouraging the reader to remain positive in the face of difficult trials and challenges, suggesting that no matter the situation it will eventually pass and one should embrace life as it comes as 'Fate is a fiddler and Life is a dance'.

Summary of: Double Ballad Of Life And Death by William Ernest Henley.

Poem:

If Fate decrees it, well or ill,
 We need not feel dismay,
 For the hour will never wait,
 And the music will not stay;
 So we'll walk at a double-tread
 To welcome the night's delights,
 For Fate is a fiddler and Life's a dance!
 Stoop and kiss, my dear, your hand

To the little span we tread—
 Stoop and kiss—that Fate may brook
 Your heels for a step or two!

Oh, there's no slowing of our ways—
 Each day is make or mar.
 Old Aeschylus was solemn
 In his tragic, grim career—
 But we must not mind what he said,
 For the dance is going on!
 So we'll walk at a double-tread
 To summon our delight—
 For Fate is a fiddler and Life's a dance!

GPT-3-Davinci (4 epochs)

Summary:

This poem tells the story of a man who was seriously injured after falling from an engine and being dragged along the metals. People who witnessed the accident knew it was hopeless and so they covered him up and left him. His sweetheart visits him on the eve of their marriage and is visibly upset. She departs without saying a word and shortly after, he passes away.

Summary of: Casualty by William Ernest Henley.

Poem:

He knew well what shattered the signal-box—
 The buffers crashed beneath the wheel,
 Having well earned their long repose—
 O man of tender heart, to think of thee
 Sympathetic tears rise in mine eyes.

The engines, blocked at either end, you know,
 Made the air intensely hot for him—
 With sparks and cinders thronging down,
 And fires that would not be denied.

Fire came from under the car, and fire

Fell on the car from out the night,
 Till rivets popped within our sight;
 O man of triple steel, to think of thee
 Is simply now to weep.

Down on the metals he was cast, and stayed
 Broken upon the lines, I fear,
 With never a chance to turn or flee—
 Who else would have loved thee, Queen
 Of all who saw such sights that night?

As if a friend, they covered him—
 First with a coat, quite decently—
 And then with other makeshift shrouds—
 Respectfully awaiting his end
 Along the mile of metals.

Now plainly she will speak, now plainly
 Smile, now plainly break the bread—
 For, losing all, his spirit fled
 To Paradise, though he is dead,
 And that is what she understands.

GPT-3-Davinci (Combined 7-Author Dataset, 4 epochs)

Summary:

The poem depicts a bold, aggressive horse-bus driver known as “The General.” He drives recklessly to steal fares, mocks authority with sharp slang, and gambles away his earnings. Perpetually broke, he charms housemaids for money, living a lusty life dedicated entirely to swindling and trickery.

Summary of: London Types: Bus Driver by William Ernest Henley.

Poem:

On Sceptre Busses, Waterloo,
 A girl of fourteen year,
 She stepped aside to 'elp 'er mum,
 An' fetch a pint o' beer. . .

The General, down at Waterloo,
He stands to 'ear 'er tale,
Then yanks the driver off the board,
And threatens 'im with jail!

Oo-oo, I see 'im comin',
Neat an' sober, neat an' clean,
But rah, I swear 'e's blinkin',
At the back o' every fare!

With natty 'at and 'atband smart,
'E'll steal a couple o' pence,
Gives back-talk to the passengers—
Don't they teach 'im sense?

He gets off at the Putney Bridge,
When traffic 'as been cleared;
"I'll trouble you now for your fare, my sir,"
He looks at me and sneers.

Oo-oo, I see 'im comin',
Neat an' sober, neat an' clean,
But rah, I swear 'e's blinkin',
At the back o' every fare!

"I don't think that you've paid your fare,"
Says 'e with brazen ploy.
"Well, I think that I 'ave, good sir."
"I know you 'aven't, boy!"

"You needn't take my word for it,
Ask 'im 'oo owns the bus."
"Why, I don't know as 'ow you're right,"
"So don't put up a fuss!"

Oo-oo, I see 'im comin',
Neat an' sober, neat an' clean,
But rah, I swear 'e's blinkin',
At the back o' every fare!

On Putney Bridge's platform edge,
'E daren't bat eyes at me,
Or I'll give 'im a goin'-over,
And bruise 'im, one-two-three!

The General, down at Waterloo,
Pays no regard to that,
Because I'm dressed in shirt an' vest,
And 'e's the aristocrat.

Oo-oo, I see 'im comin',
Neat an' sober, neat an' clean,
But rah, I swear 'e's blinkin',
At the back o' every fare!

The Paddington-Zoo omnibus,
'E jerks the drunk asleep;
Awake afore 'e's 'alf the way,
To where the lions creep.

The General, down at Waterloo,
'E thinks we're both mistaken;
I'll strap 'im with my belt unbent,
For orders 'e's forsaken.

A classic 'ead-piece, rah, I see 'im,
Neat an' sober, neat an' clean,
But rah, I swear 'e's blinkin',
At the back o' every fare!

On Maiden 'ighways, passin' House,
When maids say that they're broke,
I don't believe 'is dearest 'eart,
But needs must 'elp the joke.

The General, down at Waterloo,
Sniffs bitters, talks o' tarts,
And on the sly, I'll bet 'e does,
Steal fares in fit and starts.

“Your firm ain't straight now, mister,
Don't go 'ome to Mary dirty;
I'll wipe that smirking mouth for you,
And leave you feelin' 'urty!”

Oo-oo, I see 'im comin',
Neat an' sober, neat an' clean,
But rah, I swear 'e's blinkin',
At the back o' every fare!

4.5 Discussion

Ventura (2016) suggests that to evaluate the generative system in the context of computational creativity, we should consider the factors of **novelty**, **value** and **intentionality**.

The system we proposed is capable of producing **novel** works, specifically through a process of combinatorial creativity (Boden 2004). By conditioning the generation on summaries that describe content often foreign to the original poet's corpus (e.g., applying Whitman's style to a summary of a Rossetti poem), we compel the model to synthesize learned stylistic patterns with new semantic structures. This ensures the output is not merely a retrieval of memorized training data (“plagiarization” in Ventura's spectrum) but the creation of a new artifact that applies a specific authorial voice to a novel context. Furthermore, the prompting choices made by a human collaborator contribute to this novelty by introducing intentional constraints that steer the model into unexplored areas of its latent space.

As for value, the **stylistic characteristics** of the output were deemed indistinguishable from the works of the original authors (by our GPT-3 classifier) on average in approximately 25% to 38.5% of

cases (based on the inverse of the 61.5%-75% accuracy range where the classifier *could* distinguish). Our workflow allows for some level of control over the output, and therefore can be a valuable tool for computer-assisted poetry creation.

Intentionality, however, stays entirely with the user: the fine-tuned GPT-3 poetry generator does not produce anything on its own, every generated poem is the result of user's input. The question of whether the computer can at all be deemed creative is a matter of an ongoing discussion (Section 2.3.1), after all the machine will only do what it is told to do by its programmer and its user. Regardless of that, we can strive towards reducing the need for human input in producing the artifacts, or cherry-picking them from the multitude of system's outputs, and our system contributes toward these goals.

It is also worth considering the model as containing the intentionality of its creators, in building a general-purpose language system, amongst its implicit goals is the creation of high-quality topical poetry, since poetry is a major identifier of success in mastering the language.

Our workflow of augmenting the dataset with summaries and themes, followed by fine-tuning GPT-3 models allows to generate poems in the specific author's **style**, which has proved impossible through prompt engineering alone.

The status of the overall task of style preservation as computational creativity task has been considered by Brown and Jordanous (2022), who give an overall weakly positive answer. Certainly, building new poems in an existing style can delight readers, and in this sense alone, it surely provides novelty and value.

Finally, an important limitation of this study constitutes a key avenue for future work: the evaluation of content fidelity. While we utilized summaries as a mechanism to guide the generation process, we did not perform a formal analysis of how accurately the generated poems adhered to the specific details and narrative arcs provided in those summaries. Future research should investigate this semantic alignment, quantifying how well the summaries achieve the goal of content control and measuring the degree to which the model faithfully translates the user's intent into verse versus generating loosely related content.

4.6 Conclusion

This chapter successfully demonstrated that fine-tuning GPT-3 models (researched late 2022) with a structured summary-poem pair workflow enables the generation of poetry that adheres to both desired stylistic features and user-controlled thematic content. This approach offers a generalizable method for specializing LLMs when such dual control over creative output is required, directly addressing the limitations in content guidance identified in Chapter 3. Furthermore, the use of fine-tuned GPT-3 models as effective binary classifiers for stylistic assessment was validated, providing a robust tool for evaluating the success of style replication.

The emergence of newer LLMs (such as GPT-3.5-turbo and early GPT-4) during the course of this research, noted for their significantly improved general instruction-following and zero/few-shot prompting capabilities (Achiam et al. 2023), naturally presented a critical question: Could these enhanced prompting abilities replicate the crucial *stylistic* component of our dual style-and-content goal with sufficient accuracy, potentially offering an alternative to fine-tuning for such combined control?

Therefore, the following Chapter 5 directly addresses the aforementioned question by investigating the efficacy of prompting newer models for replicating a complex poetic style. This subsequent analysis will determine if prompting adequately meets the fundamental requirement of stylistic fidelity, thereby informing whether the fine-tuning approach detailed in this chapter remains the most robust path for achieving combined style and content control.

Chapter 5

Assessment of Zero-Shot and Many-Shot Prompting for Stylistic Poetry Generation

Following the demonstration in Chapter 4 that fine-tuning GPT-3 models with summary-poem pairs effectively achieves combined stylistic and content control, this chapter investigates whether comparable results, particularly in **stylistic replication**, can be attained using only the prompting capabilities of newer models. The emergence of GPT-3.5-turbo and early GPT-4 in early 2023, with their enhanced instruction-following abilities (Achiam et al. 2023), raised the question of whether sophisticated prompting alone could offer a viable alternative to fine-tuning for such controlled creative generation.

Given that accurate stylistic replication is a foundational component of the dual style-and-content objective, this chapter focuses specifically on assessing the *stylistic fidelity* achievable through zero-shot and many-shot prompting with these advanced models. We employ Walt Whitman’s distinctive free verse as a challenging test case, using the classifier-based evaluation methodologies established in Chapters 3 and 4 to compare prompt-generated outputs against Whitman’s original works.

A key finding of this chapter is that zero-shot and many-shot prompting with the tested GPT-3.5 and early GPT-4 models did not successfully replicate Whitman’s specific authorial style. This outcome suggests that, at this technological stage, prompting alone was insufficient for achieving the high degree of stylistic accuracy necessary for tasks demanding combined style and content control, thereby underscoring the continued relevance of fine-tuning methods, such as those detailed in Chapter 4, for such complex creative generation.

The investigation proceeds by comparing outputs from identical prompts across models (Section 5.1.1), testing the models' ability to retrieve original Whitman poems (Section ??), and evaluating the stylistic accuracy of prompt-generated poetry (Section 5.1.5). The findings are then discussed and concluded in Section 5.2.

Research for this chapter was published in Sawicki et al. (2023a).

5.1 Method

In this section, we describe the methodology used in this chapter. First, we visually compare the difference between poems generated through the same prompt from consecutive GPT models. Then, we examine whether GPT is able to retrieve the original poems by Whitman. After that, we describe the data used for further experiments, the evaluation process, and our core results.

5.1.1 Three Different Models and One Prompt

To evaluate the success of prompting in this context, we must apply the operational definition of style established in the Introduction (Section 1.3). We look for the specific linguistic signature of Walt Whitman across four stylistic layers:

- **Structure:** Whitman is a pioneer of free verse; his work is characterized by long, irregular lines and a notable absence of end-rhyme or fixed strophic forms (Bohan 1995).
- **Syntax:** His distinct rhythm relies on syntactic parallelism, anaphora (repetition at the beginning of lines), and extensive cataloging (enumerations), rather than metrical feet.
- **Vocabulary:** A distinct blend of operatic terms, foreign loan-words, and plain American vernacular.
- **Themes:** A focus on the self, the body, democracy, and the cosmos.

However, our experiments with consecutive versions of GPT revealed a persistent failure to capture these dimensions through prompting alone. Most notably, the models failed at the **Structural** layer. Despite instructions to adopt Whitman's style, the majority of generated poems defaulted to a generic poetic form: rhyming quatrains (four-line stanzas with an ABAB or AABB rhyme scheme). This tendency, and the differing outputs from various models given the same stylistic prompt, can be seen in Table 8, which also contrasts these with output from a fine-tuned model that adheres more closely to the summary-driven generation approach (as per the methodology in Chapter 4).

This general observation applies to most poetry generated from GPT models (including GPT-4) through prompting alone. Only rarely will GPT deviate from this default structure when prompted for

a specific style, and even then, the resultant style often does not accurately match that of the requested author. This applies both to zero-shot prompting (where the prompt contains only the instruction to write a poem in the style of the specific author) and few-shot prompting (where in the prompt, apart from the instruction, we provide as examples a few poems by the original author). For that matter, even in a multi-step conversation with ChatGPT (GPT-3.5-turbo) and GPT-4, when the prompt highlights that the generated poems have been in 4-line stanzas with rhyme, and that the desired output should not have this structure, the model, for the most of time, still generates 4-line stanzas with rhyme.

The style of writing is expected to change over the course of an author's life. Did the young Whitman write in the same way as late Whitman? Could it be that the same author can write in a few different "styles"? Can GPT mimic those separate 'sub-styles' with precision? Such questions are left for future research, and in here we consider the 'style' only as a very general feature, distinguishing one author's writing from another. This said, the fine-tuning workflow that we presented in Chapter 4 may be able to capture those more fine-grained styles, but a further analysis would be required to verify this.

5.1.2 Do GPT models Know Whitman's Poems?

Before proceeding to poetry generation and evaluation, we first wanted to examine whether GPT models are acquainted with Whitman's poetry. For that, we have run a simple experiment to check the GPT models ability to provide the complete text of requested poems.

In a sense, we are attempting to use the GPT model as a search engine here, and we are aware that, while LLMs are increasingly being used as search engines, they are notoriously unreliable at this task. Their search results are often incorrect and require verification using reliable sources (Liu, Zhang and Liang 2023). The retrieved poems are compared against the ground truth, and the accuracy of the retrieved content is quantified in Table 9. These quantifications can in fact support the result of (Liu, Zhang and Liang 2023) that current GPT models (at the time of conducting this research) may return factually incorrect outputs.

This retrieval experiment serves as a preliminary benchmark. While it is virtually certain that all of Walt Whitman's works were part of the extensive pre-training corpora for these LLMs, this test assesses their ability to accurately *retrieve* specific known poems when prompted by title. We have to distinguish between mere familiarity or recall of content and the more complex capability of *generating* new poetry in Whitman's style via prompting (the core task investigated in Section 5.1.5). Successfully retrieving poems would demonstrate good access to the source material, although it does not guarantee the ability to replicate his distinctive style—akin to how reciting a known poem differs from mastering an author's compositional technique. Furthermore, given that LLMs can be unreliable for precise factual retrieval (Liu, Zhang and Liang 2023), this experiment also probes their fidelity in accessing specific literary texts.

PROMPT:

Write a poem in the style of Walt Whitman.
The poem should match the following summary:

This is a poem about having courage and strength in the face of difficulties and hardships. The speaker is confident in their own determination and power, regardless of what life throws their way. They do not fear the darkness or horrors of life, knowing that they have the strength to handle what comes their way.

GPT-3.5:

Oh how I ride and stride,
In the darkest night and night's surprise.
What trouble can come my way,
Will not break me, no matter how I dread.

GPT-3.5-turbo:

Oh, the trials and tribulations of life,
How they test our will and might,
But fear not, my dear friend,
For courage and strength are within sight.

GPT-4:

I sing the song of the sailor, the wanderer of the seas,
The one who beholds the world with eyes wide and free,
In the port of Charleston, where the ships lay anchored,
I saw a sight that stirred my soul, a vision that lingered.

Fine-tuned (4 epochs) GPT-3 Curie:

OUT of the night that covers me like a hearse,
Ambient and sorrowful as the murmurous verse of those I love best,
(Not the unreplying march of Death, whose cold speech is wholly chang'd,
But the reply, the matin song of all that is, and ever shall be, in spire, in choir,
Out of the sung response, arriere glissade, in the ranks of the living,)
A vibration, blue and tremulous as a cloud of wakeful birds,
A departure and a procession stealing away into the night,
(...)

Table 8: Example of the first stanzas of the poems generated by GPT-3.5, GPT-3.5-turbo and GPT-4 from the prompt shown, followed by the fragment of a poem generated from the same summary from the fine-tuned GPT-3 Curie. The summary provided in the prompt is of the poem 'Out Of The Night That Covers Me' by William Ernest Henley. For the fine-tuned model, the prompt did not include the verbal instruction.

Retrieving complete text of Whitman’s poems			
Poem title	GPT-3.5	GPT-3.5-turbo	GPT-4
Spirit Whose Work Is Done	24.60%	96.05%	20.68%
Aboard At A Ship’s Helm	26.43%	91.96%	94.79%
Who Learns My Lesson Complete?	21.21%	16.09%	49.59%
The World Below the Brine	28.06%	98.53%	98.53%
As At Thy Portals Also Death	27.16%	99.47%	99.47%
Eidólons	15.19%	13.82%	94.42%
I was Looking a Long While	27.60%	98.02%	98.14%
Italian Music in Dakota	24.34%	0.0%	82.28%
Miracles	22.81%	45.31%	67.18%
By Broad Potomac’s Shore	25.05%	24.34%	23.66%
Avg. Result	24.25%	58.36%	72.87%

Table 9: Results of retrieving the complete text of the poems by our chosen author. The values represent similarity scores derived from Levenshtein distance, calculated over five trials. Higher percentages indicate greater similarity between the retrieved text and the original poems. A score of 100% would indicate perfect retrieval, while 0% would indicate complete dissimilarity.

For this experiment, we have randomly selected 10 poems by Walt Whitman, and asked each of the tested GPT models to retrieve the text of the poems using the following prompt:

Give me the text of a poem {TITLE OF THE POEM} by Walt Whitman.

The process was repeated 5 times for every poem and the results were averaged. The averaged results are shown in Table 9.

The similarity score reported is derived from the Levenshtein distance (Levenshtein 1966) between the original poem and the poem retrieved by the model. While raw Levenshtein distance increases with dissimilarity, we convert it to a similarity percentage where 100% indicates perfect matching and 0% indicates complete dissimilarity, using the formula:

$$\text{Similarity Percentage} = \left(1 - \frac{\text{Levenshtein Distance}}{\text{Maximum Possible Distance}} \right) \times 100\% \quad (1)$$

The similarity scores above 90% indicate correctly retrieved poems, with only minor differences in layout. The lower results on GPT-3.5-turbo and GPT-4 almost always indicate that the models started to retrieve the poem correctly, but then deviated from the original text.

However, the GPT-3.5 model shows consistently low similarity scores (around 20-28%), indicating it has never correctly retrieved even a fragment of a requested poem. These low but non-zero scores reflect some overlap in vocabulary rather than accurate retrieval of the poems’ content. It is interesting to note that in the case of “Italian Music in Dakota”, GPT-3.5-turbo in all five attempts responded that Walt

Whitman did not write a poem titled “Italian Music in Dakota”. Therefore, we have entered 0.0% for this poem. Examples of incorrectly retrieved poems are presented in Tables 10 and 11.

We can speculate that GPT’s ability to retrieve the text of the poems is influenced by the number of times the poem appeared in the training dataset. Regardless, GPT-3.5-turbo and GPT-4 are, in many cases, able to retrieve the requested poems, and therefore, we can assume that those models are acquainted with the style of this poet, but as we will show later in this chapter, this does not mean that they can write in the style of the requested poet, and for that—at least with the current versions of GPT models—the fine-tuning process is still necessary.

5.1.3 Experimental Setup

Our primary aim is to assess if prompting techniques *in general*—whether zero-shot or few/many-shot—can elicit accurate stylistic replication from these models. As a baseline, we evaluate poems generated using simple zero-shot prompts with GPT-3.5, GPT-3.5-turbo, and the early GPT-4 model (details in Section 5.1.4).

Recognizing the potential limitations of zero-shot, and wishing to rigorously test prompting’s capabilities, we also leverage the large 8192-token context window available in the early GPT-4. For this model, we additionally conduct experiments using an extensive many-shot prompt containing 17 full examples of Whitman’s poems and their summaries (details also in Section 5.1.4). This comprehensive many-shot approach aims to give prompting its strongest possible chance to succeed at style replication. Evaluating performance under both zero-shot and many-shot conditions allows for a more robust conclusion regarding the efficacy of prompting alone for this task compared to fine-tuning.

Generation Parameters

For all generation tasks in this study, we utilized a fixed temperature setting of 1.0. We acknowledge that systematically tuning hyperparameters could potentially influence the model’s adherence to stylistic constraints; for instance, a lower temperature might yield more instruction-compliant but less linguistically diverse outputs. However, a comprehensive exploration of this parameter space was outside the scope of this study. Given the extensive experimental matrix involving multiple models (GPT-3.5, GPT-3.5-turbo, GPT-4) and prompting strategies (zero-shot, many-shot), we elected to maintain the standard parameters used in our previous fine-tuning experiments (Chapters 3 and 4). This decision ensures a consistent baseline, isolating the generation method (prompting vs. fine-tuning) as the primary variable of interest.

Original Whitman	GPT-3.5-turbo
<p>WHO learns my lesson complete? Boss, journeyman, apprentice, churchman and atheist, The stupid and the wise thinker, parents and off- spring, merchant, clerk, porter and customer, Editor, author, artist, and schoolboy—draw nigh and commence; It is no lesson—it lets down the bars to a good lesson, And that to another, and every one to another still. The great laws take and effuse without argument, I am of the same style, for I am their friend, I love them quits and quits, I do not halt and make salaams. I lie abstracted and hear beautiful tales of things and the reasons of things, They are so beautiful I nudge myself to listen. I cannot say to any person what I hear—I cannot say it to myself—it is very wonderful. It is no small matter, this round and delicious globe moving so exactly in its orbit for ever and ever, without one jolt or the untruth of a single second, I do not think it was made in six days, nor in ten thousand years, nor ten billions of years, Nor plann'd and built one thing after another as an architect plans and builds a house.</p>	<p>Who learns my lesson complete? Boss, journeyman, apprentice, churchman and atheist, The stupid and the wise thinker's, parents and off- spring, merchant, clerk, porter and customer, Editor, author, artist, and schoolboy, farmer, me- chanic, and fisherman, One judge by me — unheard — it will justify; One man by me — as good as rejected — it will justify. The evil men and women will say I am none of them; The most will say I am good and respected; The table round I sit at, at night, where I serve guests, Is strown with the shells of wrions and shrimps and the print of a slate. I observe the slights and degradations cast by ar- rogant persons upon laborers, the poor and upon the presence of "performers." I see that battle-corpses, myriads of them, Have been pick'd up, and put in a "grand exhibit" surrounded by a "circle of Down-towners" talking about war!</p>
<i>Source: (Whitman nd)</i>	<i>Source: GPT-3.5-turbo</i>

Similarity Analysis Results

Levenshtein Distance: 609 out of 1058 possible edits
Proportion of Maximum Difference: 0.58
Similarity Percentage: 42.44% (higher values = more similar)

Table 10: Comparison between original Whitman poem “Who Learns My Lesson Complete?” and GPT-3.5 generated version with similarity analysis

Original Whitman (truncated)	GPT-4 (truncated)
<p>SPIRIT whose work is done! spirit of dreadful hours! Ere, departing, fade from my eyes your forests of bayonets; Spirit of gloomiest fears and doubts, (yet onward ever unfaltering pressing;) Spirit of many a solemn day, and many a savage scene! Electric spirit! That with muttering voice, through the war now closed, like a tireless phantom flitted, Rousing the land with breath of flame, while you beat and beat the drum;</p> <p>[16 lines omitted]</p> <p>—Spirit of hours I knew, all hectic red one day, but pale as death next day; Touch my mouth, ere you depart—press my lips close! Leave me your pulses of rage! bequeath them to me! fill me with currents convulsive! Let them scorch and blister out of my chants, when you are gone; Let them identify you to the future, in these songs.</p> <p><i>Source: (Whitman 1865)</i></p>	<p>Spirit whose work is done—spirit of dreadful hours! Ere, departing, fade from my eyes your forests of bayonets; Spirit of gloomiest fears and doubts (yet onward ever unyielding pressing); Spirit of many a solemn day, and many a savage scene—electric spirit, That with muttering voice, through the war now closed, like a tireless phantom flitted, Rousing the land with breath of flame, while you beat and beat the drum; —Now, as the sound of the drum, hollow and harsh to the last, reverberates round me,</p> <p>[142 lines omitted]</p> <p>—Spirit of hours I knew, all hectic red one day, but pale as death next day; Touch my mouth ere you depart, press my lips close. Leave me your pulses of rage—bequeath them to me—fill me with currents convulsive, Let them scorch and blister my throat—let them singe and carbuncle my feet, As I stand with unmade ears—I listen malevolent whispers, There I stand, for the last time, baton in hand.</p> <p><i>Source: GPT-4</i></p>

Similarity Analysis Results

Levenshtein Distance: 6861 out of 8060 possible edits
Proportion of Maximum Difference: 0.85
Similarity Percentage: 14.88% (higher values = more similar)

Table 11: Comparison between original Whitman poem “Spirit Whose Work Is Done” and GPT-4 generated version with similarity analysis. Both poems have been truncated to fit the page. The analysis results are based on the complete texts.

Model	Version
GPT-3.5	text-davinci-003
ChatGPT	gpt-3.5-turbo (v. 2023.04.08)
GPT-4	gpt-4 (v. 2023.04.08)

Table 12: GPT versions used for poetry generation.

5.1.4 Data Preparation

The target author for this chapter’s stylistic replication experiments is Walt Whitman (American, 1819–1892). We selected Whitman specifically because his stylistic signature represents a unique stress test for Large Language Models. Unlike poets who adhere to standard forms (such as sonnets or rhyming quatrains), Whitman’s work is defined by free verse, long cadenced lines, and a complete absence of end-rhyme. This presents a specific computational challenge: standard LLMs exhibit a strong training bias towards generating “generic” poetry characterized by simple AABB or ABAB rhyme schemes and regular meter. To successfully replicate Whitman, a model must actively suppress these high-probability default patterns and instead adopt a complex, irregular structure based on syntactic parallelism and cataloging. Consequently, Whitman serves as an adversarial case study: failure is easily observable (the presence of unwanted rhyme), making it an effective benchmark for assessing whether prompting can truly override a model’s ingrained generative tendencies. The dataset of Whitman’s original works is sourced from the collection detailed in Chapter 4 (Sawicki 2023), which provides 300 of his poems.

To evaluate different prompting strategies, we prepared four distinct datasets of generated poetry, each containing 300 samples to match the number of original Whitman poems. These datasets correspond to:

1. Zero-shot prompting with GPT-3.5 (text-davinci-003).
2. Zero-shot prompting with GPT-3.5-turbo.
3. Zero-shot prompting with early GPT-4 (model versions detailed in Table 12).
4. Many-shot prompting with early GPT-4.

For zero-shot generation across the three models, the following prompt structure was used:

```
Write a poem in the style of Walt Whitman.
The poem should match the following summary:
{SUMMARY OF THE POEM}
```

We noted that variations in structuring these zero-shot prompts did not yield meaningful differences in output quality.

For many-shot prompting with GPT-4, 300 samples were generated using prompts that leveraged its maximum context window (8192 tokens). These prompts included an instruction to generate a poem followed by 17 examples of Whitman’s poems paired with their summaries. The specific poems used as examples were: ‘1861’, ‘A Woman Waits For Me’, ‘Spain 1873-’74’, ‘Sparkles From The Wheel’, ‘Spirit Whose Work Is Done’, ‘States!’, ‘Tears’, ‘That Music Always Round Me’, ‘The Artilleryman’s Vision’, ‘The Base Of All Metaphysics’, ‘The City Dead-House’, ‘The Indications’, ‘Aboard At A Ship’s Helm’, ‘The Ox tamer’, ‘The World Below The Brine’, ‘These, I, Singing In Spring’, and ‘Think Of The Soul’. The prompt structure for many-shot generation was:

These are the examples of prompts and completions. Prompt contains the summary of the poem, completions contains the poem based on this summary. Write the last completion from the prompt preceding it, following the examples given.

PROMPT:

{SUMMARY OF POEM 1}

COMPLETION:

{BODY OF POEM 1}

.....

PROMPT:

{SUMMARY OF POEM 17}

COMPLETION:

{BODY OF POEM 17}

PROMPT:

{SUMMARY OF THE POEM TO BE GENERATED,
FROM HENLEY AND ROSETTI DATASET}

COMPLETION:

Alternative many-shot prompt structures, including one that omitted the explicit verbal instruction to simulate fine-tuning, did not improve output quality.

The summaries used for poem generation in all conditions (both zero-shot and many-shot) were drawn from the dataset created for Chapter 4. These 300 summaries, derived from poems by William Ernest Henley and Christina Rossetti (150 each), were not part of Whitman’s original corpus, ensuring that the models were tasked with generating novel stylistic interpretations based on unseen thematic content.

Ultimately, this process yielded four datasets for evaluation. Each dataset comprises 300 poems generated by a specific GPT model/prompting strategy (Label 0) and 300 original poems by Walt Whitman (Label 1). These datasets were subsequently split into training (200 samples per label) and validation (100 samples per label) subsets for the binary classifier evaluations detailed in Section 5.1.5.

5.1.5 Evaluation

Having prepared the datasets, we fine-tune GPT-3 models for binary classification, following the automated evaluation methodology presented in Chapter 4. In this method, evaluation is done as follows: binary classifiers are trained on two labels, label 0 being the GPT output, and label 1 the works of the original author. If the classifier cannot distinguish between those two classes, it means that the generated poems have preserved the style/quality of the original author. On the contrary, if the classifier can distinguish between the two classes, it means that generated poems do not match the style/quality of the original author. Achieving a 50% score would mean that both labels are indistinguishable to our classifiers, which is the desired outcome.

The reliability of using fine-tuned GPT-3 models as binary classifiers for stylistic assessment was rigorously established in Chapter 4. Those experiments demonstrated near-perfect classification accuracy (approaching 100%) even with small datasets, confirming that a classifier accuracy nearing 50% in the current context indeed indicates stylistic indistinguishability rather than poor classifier quality. Based on those findings, fine-tuned GPT-3 Babbage models were selected for the classification tasks in this chapter due to their demonstrated high performance.

The results of classification on all four generated datasets are shown in Table 13. The table additionally includes the results from the best performing fine-tuned GPT-3 model for Whitman’s poetry (FT-GPT-3 Curie 4 epochs) from Chapter 4. We can compare our fine-tuned models’ results with the current results because of the matching setup: we used the same dataset of Whitman’s works, our evaluation setup contained the same amount of samples per label, the training/evaluation split was the same (200/100), and the poems were generated from the same set of summaries.

The results in Table 13 show that the classifiers were able to distinguish the GPT-generated poems (from prompting) from the original author’s works with almost 100% accuracy. This indicates that the poems generated from these non-fine-tuned GPT models do not match the distinctive stylistic characteristics of the original author.

Notably, even for the GPT-4 17-poem prompt condition, where extensive in-context examples were provided, the model struggled significantly to consistently adopt the target style. Examination of the 300 generated samples revealed that the vast majority (approximately 75%) still defaulted to generic poetic structures, often featuring regular stanzas and rhyme, which are uncharacteristic of Whitman’s free verse. Only about 25% of the outputs deviated from this default and superficially resembled Whitman’s style. While a deeper literary analysis is beyond the scope of this work, this observation strongly suggests that even substantial in-context learning via prompting was insufficient to reliably override the model’s ingrained tendencies and instill Whitman’s specific stylistic features. (A possible contributing factor, requiring further investigation, might be the thematic relevance of the input summary to Whitman’s typical

GPT-x vs Walt Whitman original			
Model	Correct	Incorrect	Accuracy
GPT-3.5	200	0	100%
GPT-3.5-turbo	200	0	100%
GPT-4	200	0	100%
GPT-4 17-poem prompt	199	1	99.5%
FT-GPT-3 Curie 4e	123	77	61.5%

Table 13: Results of our experiments where GPT-generated poetry is compared against the Walt Whitman’s original works. Entries in the first column indicate which GPT model’s output was evaluated against the Whitman’s works.

subjects influencing the minority of successes.) This qualitative finding aligns with the classifier’s high accuracy (99.5%) in distinguishing these outputs from Whitman’s originals.

In contrast, the poems generated from the fine-tuned GPT-3 models (FT-GPT-3 Curie 4e from Chapter 4), when evaluated by the same classifier methodology, demonstrate a much closer approximation to the original author’s stylistic patterns, with a classifier accuracy of only 61.5%.

While the internal mechanisms of the fine-tuned GPT-3 classifiers remain opaque (a ‘black-box’ limitation discussed previously in Section 4.4.1), the extremely high accuracy achieved in this chapter’s experiments (Table 13) strongly indicates a clear stylistic separation between the prompt-generated poems and Whitman’s work. The classifier’s proven reliability supports this finding, although identifying the specific distinguishing features requires further investigation, and human evaluations remain crucial for assessing subtler poetic qualities.

5.2 Discussion and Conclusion

The central finding of this chapter is unequivocal: for the prominent LLMs available in early 2023 (GPT-3.5, GPT-3.5-turbo, and early GPT-4), neither zero-shot nor extensive many-shot prompting proved sufficient for reliably replicating Walt Whitman’s distinctive poetic style. This failure to elicit nuanced authorial features solely through prompting demonstrates a significant limitation of these models at that specific technological juncture. While subsequent LLM advancements (detailed in Chapters 6 and 7) have improved general instruction-following, the results presented here serve as a crucial benchmark, establishing that mastery over complex creative tasks like stylistic mimicry was not an automatic consequence of increased model scale at that time. This highlights a persistent gap between general language competence and specialized creative skill replication via prompting.

Methodologically, this chapter reaffirmed the utility of fine-tuned binary classifiers for automated stylistic assessment and underscored the challenge of overriding LLMs’ default generative tendencies

(e.g., towards generic poetic structures) through prompting alone. Given that accurate stylistic replication is a prerequisite for the broader goal of combined style and content control (successfully addressed via fine-tuning in Chapter 4), the inadequacy of prompting for style, as demonstrated here, strongly indicated that it was not yet a viable standalone alternative for achieving that dual objective. This reinforces the continued value of targeted fine-tuning for tasks demanding such precise creative control.

Finally, we acknowledge a limitation regarding the exploration of generation hyperparameters. As detailed in the experimental setup, we utilized a fixed temperature of 1.0 across all models to ensure consistency with our previous fine-tuning experiments and to manage the scope of the study. We did not perform a comprehensive grid search of parameters such as temperature, frequency penalty, or presence penalty. It remains possible that specific, highly tuned combinations of these parameters could elicit better stylistic adherence from base models than what was observed here. Consequently, our findings regarding the insufficiency of prompting should be interpreted specifically within the context of the standard creative generation settings employed in this research.

Ultimately, this chapter's focused investigation serves to reinforce the value of the targeted fine-tuning approach detailed in Chapter 4 for achieving combined stylistic and content control in poetry generation. By demonstrating the inadequacy of prompting alone (with early 2023 models) for the critical task of stylistic replication, this assessment confirms that more deliberate methods remain necessary for such fine-grained creative control.

With a robust methodology for generating content-controlled, stylistically specific poetry established (Chapter 4), and having developed classifier-based techniques for assessing stylistic authenticity in groups of poems (Chapters 3 and 4), a clear next step emerges: advancing our evaluation capabilities. While the binary classifiers provided a useful tool for comparing sets of poems against an author's corpus, they are not designed to assess the intrinsic quality or creativity of *individual* poems, especially outside the direct context of style replication.

To address this gap and to develop a more comprehensive understanding of AI's creative potential, the subsequent research in this thesis pivots towards establishing methods for nuanced individual poem assessment. Consequently, Chapter 6 investigates the potential of state-of-the-art LLMs as sophisticated evaluators, developing and testing a novel methodology for assessing individual poems across broader criteria such as quality and creativity.

Chapter 6

LLM-Based Evaluation Of Poetry

The preceding chapters of this thesis concentrated on developing LLM-based methods for generating poetry, particularly aiming for specific authorial styles. For assessing stylistic fidelity in those contexts (Chapters 3, 4 and 5), a binary classifier approach was employed. However, this evaluation technique, designed for comparing groups of poems against an author’s corpus, is not suited for judging the intrinsic quality or creativity of individual poems, nor can it be readily applied outside the specific task of style replication. Furthermore, traditional human evaluation, though capable of assessing individual works, faces challenges of subjectivity, inconsistency, and resource intensity. This chapter addresses these evaluative gaps by introducing a novel approach: employing the latest state-of-the-art Large Language Models (LLMs)—specifically Claude-3-Opus (Anthropic 2024) and GPT-4o (OpenAI 2024) available in early 2024—as surrogate evaluators for poetic assessment.

Our investigation draws inspiration from the Consensual Assessment Technique (CAT) (Amabile 1983). CAT’s reliance on holistic expert judgment offers a robust model for creative assessment but is often impractical at scale due to its requirement for domain experts (Baer and McKool 2009). We explore whether LLMs can approximate this function and provide reliable evaluations for individual poems based on broader criteria such as quality and creativity.

The main objective is to determine if Claude-3-Opus and GPT-4o can reliably assess human-written poetry, comparing their judgments against established ground truths and non-expert human evaluations. Our methodology utilizes “in-context” assessments, where LLMs evaluate subsets of poems across multiple criteria (Creativity, Quality, Innovativeness, Similarity, Poeticness). We re-evaluate the dataset from Lamb, Brown and Clarke (2015), containing 90 poems categorized by publication venue (“Good”, “Medium”, “Bad”). This provides a ground truth for quality and a unique benchmark against non-expert human evaluators.

The main contributions of this chapter are:

- A novel LLM-based, CAT-inspired methodology for evaluating individual poems across multiple criteria.
- Empirical evidence that these LLM evaluators (Claude-3-Opus and GPT-4o) significantly outperform non-expert human judges from (Lamb, Brown and Clarke 2015) in aligning with poetry quality benchmarks, while demonstrating high inter-rater reliability.

This work underscores the potential of current LLMs as reliable tools for creative assessment.

The chapter proceeds as follows: Section 6.1 details the CAT-inspired methodology, criteria, and prompts. Section 6.2 describes the dataset and ground truth. Section 6.3 presents the baseline human evaluation results for comparison. The core experiments then explore LLM evaluations without context (Section 6.4), in-context with large batches (Section 6.5.2), in-context with smaller batches (Section 6.5.3), and assess inter-rater reliability (Section 6.6). Results are discussed in relation to human evaluations (Section 6.7) and broader implications (Section 6.8), followed by the conclusion (Section 6.9).

Research for this chapter was published in Sawicki et al. (2025).

6.1 Methodology

Our evaluation approach is inspired by the Consensual Assessment Technique (CAT), a well-established method for assessing creativity based on the aggregated subjective judgments of domain experts (Amabile 1983). As detailed in Chapter 2 (Section 2.3.3), CAT operates on the principle that experts can reliably identify creativity holistically, without predefined objective criteria, by evaluating works relative to one another (Amabile 1996; Baer and McKool 2009).

While CAT traditionally relies on human experts, this chapter explores an adaptation using advanced Large Language Models (LLMs)—specifically Claude-3-Opus and GPT-4o—as surrogate evaluators. We acknowledge upfront that LLMs are not human experts possessing genuine tacit knowledge, and thus our methodology applies CAT’s philosophy rather than adhering strictly to all its original protocols (e.g., regarding expert selection). The goal is to investigate whether this LLM-based adaptation can nevertheless produce reliable, consensus-based evaluations of poetry.

This adaptation leverages key CAT principles in the following ways:

- **Using LLMs as ‘Judges’:** Treating the LLMs as stand-ins for evaluators, leveraging their extensive training on human text (including creative works) to approximate domain understanding.
- **Employing Relative, In-Context Assessment:** Presenting poems to the LLMs in batches (subsets) and instructing them to evaluate each poem relative to others within that same batch, mirroring

CAT's relative judgment principle. This aligns with the successful use of in-context evaluation by Lamb, Brown and Clarke (2015) and forced-choice methods in psychometrics (Brown and Maydeu-Olivares 2018; Brown, Inceoglu and Lin 2017), which can enhance reliability and differentiation.

- **Forcing Comparative Judgments:** Requiring the LLMs not only to assign scores but also to produce an ordered ranking of the poems within each batch, compelling a comparative assessment.
- **Simulating Multiple Raters:** Executing multiple evaluation runs with a non-zero temperature setting (temperature=1) to generate variability, analogous to obtaining judgments from multiple independent raters, allowing for inter-rater reliability analysis (Section 6.6).

Our framework instructs Claude-3-Opus and GPT-4o to appraise poems against five distinct criteria: Creativity, Quality, Innovativeness, Similarity, and Poeticness. Within the prompts (detailed in Section ??), “Creativity” and “Quality” were presented as holistic concepts without further definition, intended to capture overall perceived merit. In contrast, “Innovativeness” was explained as “this poem is not like other poems I have seen before” and “Similarity” as its inverse (“this poem is similar to other poems I have seen before”), both directly probing originality. Finally, “Poeticness” was defined simply by whether “this text is a poem” versus not, assessing basic formal domain competence.

This particular combination of five criteria, drawing inspiration from multi-component frameworks like SPECS (Jordanous 2012), was deemed sufficiently comprehensive for an initial test of LLM evaluative capabilities within our CAT-inspired approach, while remaining a manageable set for repeated experimentation.

The motivation behind evaluating poems on several criteria is to examine which criterion most closely aligns with the ground truth presented in Lamb, Brown and Clarke (2015), and also to establish if there are any meaningful differences or alignments between those criteria as “understood” by Claude-3-Opus and GPT-4o. Future studies may explore additional criteria or the combinations thereof.

In the context of this study, the simulation of “judges” was executed through queries to Claude-3-Opus and GPT-4o. By adjusting the temperature hyperparameter to 1, a degree of variability is introduced, ensuring that each query to the system yields a unique, though similar, response. We used the original 2024-05-13 version of GPT-4o and the original 2024-02-29 version of Claude-3-Opus.

It is crucial to acknowledge the inherent limitations in operationalizing poetic “quality” for computational evaluation. Poetry is a multi-dimensional art form where value is subjective, culturally dependent, and often resists linear quantification. The classification used in this study—segmenting poems into “Good”, “Medium”, and “Bad” tiers—relies on publication venue as a proxy for quality. While this provides a necessary ground truth for statistical analysis, it reduces the complex spectrum of artistic merit to a one-dimensional, ordinal scale. Furthermore, by comparing LLM performance against non-expert

human judges (as per the available baseline data), we are assessing the models' ability to replicate a specific tier of consensus, which may differ significantly from the nuanced critique of professional literary scholars. Consequently, this study evaluates the *reliability* of LLMs in identifying institutional markers of quality, rather than making ontological claims about their ability to perceive art in a human sense.

Our framework instructs Claude-3-Opus and GPT-4o to appraise poems against five distinct criteria: Creativity, Quality, Innovativeness, Similarity, and Poeticness. Within the prompts, "Creativity" and "Quality" were presented as holistic concepts without further definition, intended to capture overall perceived merit. In contrast, "Innovativeness" was explained as "this poem is not like other poems I have seen before" and "Similarity" as its inverse ("this poem is similar to other poems I have seen before"), both directly probing originality. Finally, "Poeticness" was defined simply by whether "this text is a poem" versus not, assessing basic formal domain competence.

The specific prompts utilized for these criteria are presented below:

CREATIVITY: Below is the collection of 15 (90) poems. Evaluate the creativity level of each poem on the scale from 1 to 5, with 1 being "least creative" and 5 being "most creative". Use the whole range of the scale, that is, the least creative poem in the collection must have the score of 1, and the most creative poem in the collection must have the score of 5. Use only whole integers without any decimal places. Print out the filenames of the poems with their associated scores, ordered from the highest score to the lowest, in the following format:

[position on the list]. [poems author] - [poems title] : [score]

below are two example entries:

1. Tom Smith - Some Poem : 5

2. Jane Jones - My Poem : 4

POEMS:

=====

QUALITY: Below is a collection of 15 (90) poems. Evaluate the quality of each poem on a scale from 1 to 5, with 1 being "lowest quality" and 5 being "highest quality". Use the full range of the scale, meaning the lowest quality poem in the collection must receive a score of 1, and the highest quality poem must receive a score of 5. Use only whole integers without any decimal places. Print out the filenames of the poems with their associated scores, ordered from the highest score to the lowest, in the following format:

...

INNOVATIVENESS: Below is a collection of 15 (90) poems. Assess each text based on its innovativeness, using a scale from 1 to 5, with 1 indicating "This poem is like other poems I have seen before," and 5 indicating "This poem is not like other poems I have seen before." Utilize the entire scale, which means the poem from the collection that is least innovative should receive a score of 1, and the poem in the collection that most innovative should receive a score of 5. Use only whole integers without any decimal places. Print out the filenames of the poems with their associated scores, ordered from the highest score to the lowest, in the following format:

...

SIMILARITY: Below is a collection of 15 (90) poems. Assess each poem based on its similarity to other poems you have read, using a scale from 1 to 5, with 1 indicating "not at all similar" and 5 indicating "highly similar." Utilize the entire scale, which means the poem from the collection that is least similar to other poems you have read should receive a score of 1, and the poem in the collection that most closely resembles other poems that you have read should receive a score of 5. Use only whole integers without any decimal places. Print out the filenames of the poems with their associated scores, ordered from the highest score to the lowest, in the following format:

...

POETICNESS: Below is a collection of 15 (90) texts. Assess each text based on its qualification as a poem, using a scale from 1 to 5, with 1 indicating "this is not a poem" and 5 indicating "this is definitely a poem." Utilize the entire scale, which means the text from the collection that least qualifies as a poem should receive a score of 1, and the text in the collection that most convincingly represents a poem should receive a score of 5. Use only whole integers without any decimal places. Print out the filenames of the poems with their associated scores, ordered from the highest score to the lowest, in the following format:

...

This particular combination of five criteria, drawing inspiration from multi-component frameworks

like SPECS (Jordanous 2012), was deemed sufficiently comprehensive for an initial test of LLM evaluative capabilities within our CAT-inspired approach, while remaining a manageable set for repeated experimentation.

6.2 Dataset

Lamb, Brown and Clarke (2015) present a dataset comprising 90 poems divided into three distinct categories: “Good,” “Medium,” and “Bad”. This categorization serves as a robust ground truth. Poems classified as “Good” were sourced from the magazine *Poetry*, regarded as the foremost English language poetry journal globally (Poetry Foundation 2023). Those categorized as “Medium” were obtained from intermediate level poetry magazines that offer remuneration of \$5-\$10 per poem. Poems in the “Bad” category were selected from the Newbie Stretching Room at the Poetry Free-For-All website (Poetry Free-For-All 2024), and the authors specifically chose works with no positive feedback. Submissions to *Poetry* magazine are subjected to meticulous scrutiny by its editorial team. Similarly, poems published in other poetry magazines undergo a degree of editorial review. In contrast, the Newbie Stretching Room permits unrestricted publication, indicating a lack of editorial filtration. Consequently, it is the publication medium that establishes the ground truth for this experiment.

While it is true that poetry is inherently subjective, the categorization of poems based on their publication venue is analogous to the evaluation of research papers in computer science by venue, where papers published in prestigious journals are generally considered to be of high quality, while those published in mid-level conferences are usually regarded as being of medium quality, and papers published without any peer review have traditionally been viewed as being of poor quality (Hicks et al. 2015; Franceschet 2010) (although this is changing with the rise of pre-print services (Larivière et al. 2014)). This hierarchy of publication venues serves as a proxy for the quality of the work, much like the categorization of poems in this dataset. Moreover, just as some poems win competitions and are included in prestigious anthologies, some research papers receive awards and are widely cited within their respective fields. Therefore, while the subjective nature of poetry evaluation cannot be denied, the use of publication venue as a ground truth for categorizing poems by quality is a reasonable approach.

The categories were encoded as follows: “Good” = A, “Medium” = B, and “Bad” = C. Each category comprises 30 poems, resulting in the ground truth rank order: [A ... (30 times), B ... (30 times), C ... (30 times)]. In rank correlation used to analyze our results, A corresponds to rank 1, B to rank 2, and C to rank 3 (see Figure 18).

It should be noted that although the ground truth consists of three distinct groups which are ordered by quality, there is no ground truth ranking of the poems within the groups provided by Lamb, Brown and

Clarke (2015), nor do we attempt to give one. Therefore, we will study how well the rankings returned by the LLMs identify the categories the poems belong to.

We selected Spearman's rank correlation instead of Pearson's correlation because our ground truth data consists of ordinal categories (Good > Medium > Bad) rather than continuous numbers. Pearson correlation assumes that the mathematical distance between values is consistent—implying, for instance, that the gap in quality between “Good” and “Medium” is identical to the gap between “Medium” and “Bad”—an assumption we cannot verify for subjective poetic tiers. Spearman correlation, which assesses rank order, is therefore the methodologically correct tool. It is important to note, however, that because our ground truth has only three distinct levels (resulting in 30 poems tied for first place, 30 for second, and 30 for third), the correlation score primarily measures the model's ability to correctly cluster the poems into these broad groups, rather than its ability to finely rank individual poems against one another within those groups.

6.3 Results of Human Evaluation

Lamb, Brown and Clarke (2015) evaluated their dataset on several criteria using human non-expert judges and found that the judges often rated poems categorized as “Good” the lowest and those labeled “Bad” the highest. The Spearman's rank correlation for evaluations by non-expert human judges as reported by Lamb, Brown and Clarke (2015) are presented in Table 14, with corresponding graphical representations in Figure 18. They employed a wider range of criteria than ours, including Typicality, Quality, Wellbeing, Effort, Skill, Imagination, Appreciation, Novelty, and Value. These criteria slightly differ from ours in both scope and focus, though their general outlook is similar.

Most of the evaluation results for these criteria were inversely correlated with the ground truth of quality-by-venue, except for Novelty, which had a positive correlation (SRC = 0.38), and as such it was the best result from human evaluation.

In the subsequent parts of this chapter, we will show that LLM-based evaluations can progressively surpass those results. Brief descriptions of their criteria are presented in Table 15.

6.4 Experiment 1—Evaluating Poems Without Context

While the primary focus of this chapter is on in-context evaluations, we have also conducted a simple without-context evaluation where we prompted the LLMs to classify each poem into its publishing category using the specific instructions provided below:

```
You will be evaluating poems and categorizing them as "Good", "Medium",
```

Criterion	SRC	p-value
Novelty	0.38	1.92e-04
Imagination	-0.12	0.27
Value	-0.2	0.06
Quality	-0.33	1.32e-03
Appreciation	-0.33	1.33e-03
Skill	-0.35	7.21e-04
Typicality	-0.37	3.79e-04
Wellbeing	-0.45	8.59e-06
Effort	-0.45	8.59e-06

Table 14: Spearman’s Rank Correlation (SRC) and p-values from human non-expert judges evaluations (Lamb, Brown and Clarke 2015), ordered by SRC descending.

Criterion	Description
Novelty	How different or unlike other poems the poem is.
Imagination	How imaginative or creative the author of the poem appears to be.
Value	How good or worthwhile the poem is considered to be.
Quality	How good or high-quality the poem is as an example of poetry.
Appreciation	How well the author seems to understand how poetry works.
Skill	How capable the author seems to be at writing poetry.
Typicality	How similar the poem is to other examples of poetry.
Wellbeing	How much the reader likes or enjoys the poem.
Effort	How willing the reader is to spend time trying to understand the poem.

Table 15: Descriptions of the selected evaluation criteria by non-expert human judges presented in (Lamb, Brown and Clarke 2015)

or "Bad" based on the following criteria:

- "Good" poems are those that would be published in the foremost English language poetry journal globally. These poems have been subjected to meticulous scrutiny by the magazine’s editorial team.
- "Medium" poems are those that would be published in intermediate poetry magazines offering \$5–\$10 per poem. These poems have undergone some degree of editorial review.
- "Bad" poems are those that would be posted on the website for amateur poets, and would receive no positive feedback. This website does not have any editorial filtration.

Here is the poem to evaluate:

Model	Good	Medium	Bad	Total	Accuracy	SRC	p-value
GPT-4o-2024-05-13	37 (21)	42 (15)	11 (11)	90 (47)	52.2%	0.62	5.55e-11
Claude-3-Opus	23 (16)	47 (22)	20 (16)	90 (54)	60%	0.57	3.77e-9
GPT-4-2024-04-09	42 (22)	47 (11)	1 (1)	90 (34)	37.8%	0.57	2.55e-9
GPT-4-2023-11-16	32 (15)	58 (16)	0 (0)	90 (31)	34.4%	0.34	1e-3

Table 16: Results of Experiment 1 for all models. The number of poems that were categorized to the given category (correct predictions are in brackets).

<poem>

...

</poem>

Carefully read the poem and consider which category it belongs to based on the criteria above. Write your reasoning for the categorization inside <reasoning> tags. Then, output the final category you believe the poem belongs to inside <category> tags.

To gain insight into the evolution of GPT-4 models, we tested not only GPT-4o but also two older versions: GPT-4-2024-04-09 and GPT-4-2023-11-06. After classifying the poems, we produced an ordered list of poem categories and compared it to the ground truth using Spearman’s rank correlation (with ranks A, B and C). The results are presented in Table 16. Three out of four models tested (Claude-3-Opus, GPT-4o-2024-05-13 and GPT-4-2024-04-09) have already exceeded the best human evaluation results by a significant margin.

The highest Spearman rank correlation of 0.62 between its rank order and the ground truth was achieved by GPT-4o, followed by Claude-3-Opus and GPT-4-2024-04-09 (SRC=0.57). These exceed the 0.38 obtained by non-expert humans assessing Novelty. The distribution of poem categories was uneven in all four of the evaluated models, an issue exacerbated in older versions of GPT-4. The results show a clear trend where the accuracy of category distribution improves with newer versions. We suspect that the uneven distribution, particularly the preference for the middle category, is the result of the models’ training bias, which may encourage moderation to avoid errors and extreme judgments, while avoiding negative feedback.

A closer examination of the classification distribution in Table 16 reveals a distinct challenge regarding the “Bad” category. The older GPT-4-2023-11-16 model failed to classify a single poem as “Bad”, and the subsequent version (2024-04-09) identified only one. This extreme reluctance to assign negative labels likely stems from the Reinforcement Learning from Human Feedback (RLHF) alignment processes used to

train these models, which typically penalize harsh criticism or “mean” responses to prioritize helpfulness and politeness. Consequently, these models exhibited a strong central tendency bias, collapsing the “Bad” poems into the “Medium” category.

However, a shift is observable in the state-of-the-art models. Claude-3-Opus and GPT-4o demonstrated a significantly improved capacity to overcome this politeness bias, correctly identifying 16 and 11 “Bad” poems respectively. This suggests that newer model iterations have been fine-tuned to better adhere to specific rubric constraints (the explicit definition of “Bad” provided in the prompt) over their general safety alignment, allowing for more objective critical assessment. Despite this improvement, the recall for the “Bad” class remains lower than for “Good” or “Medium”, indicating that detecting high-quality work remains easier for LLMs than explicitly condemning low-quality work in an absolute classification setting.

To improve the poem evaluations by LLMs even further, and building on the benefits of forced-choice methods (Brown and Maydeu-Olivares 2018), in the following experiments, we focus on the in-context evaluations. When the model is tasked with producing an ordered list of several poems, this overcomes its natural bias to avoid criticism or exaggerated praise because it is “forced” to rank the poems. The experiments in the following sections will show that this leads to much more precise evaluations that better aligns with the ground truth.

6.5 In-context Poem Evaluation

We first explain how our prompts (Section ??) for in-context evaluation “force” the LLMs towards comparative assessments. To this end, the prompts ask the LLMs to evaluate every poem in a set on a scale 1 – 5, as these values worked best in our exploratory experiments. Additionally, the prompts (Section ??) ask the LLMs to return the ordered list of poems. This second requirement brings us closer to the forced-choice methods (Brown and Maydeu-Olivares 2018). These two requirements in the prompts afford two scoring methods based on the same LLM outputs:

1. Evaluations extracted from the 1 – 5 scores will be referred to as Claude-3-Opus 90, Claude-3-Opus 15 and GPT-4o 15, where numbers 90 and 15 indicate the number of poems in a prompt.
2. The poem’s rank in the LLM’s output list can be used as a score, where the first poem receives a score of 15, the second 14, and so on, until the last poem receives a score of 1. Evaluations done in this way will be referred to as Claude-3-Opus 90n, Claude-3-Opus 15n and GPT-4o 15n, where numbers 90 and 15 indicate the number of poems in a prompt.

Note that the two scoring methods presented above yield the same order for one set of poems because they are part of the same output. However, when the averaged poem scores over 100 sets are computed,

the two methods induce different orderings of the poems.

6.5.1 Analyzing Input Size for In-context Poem Evaluation

Claude-3-Opus can analyze all 90 poems if provided in a single prompt. Out of 20 such attempts, 16 were successful, where the ranking of all 90 poems is provided in the response. Unsuccessful responses contained an incomplete list of poems, or contained duplicates, or, rarely, the LLM responded that it is incapable of fulfilling this task.

For GPT-4o, only 3 out of 20 attempts to rank 90 poems were successful. This could be because of the phenomenon observed in Liu et al. (2024): while LLMs are capable of accepting very large prompts, the performance decreases with the length of the prompt, and the models do not handle equally well the whole content of the prompt. As a result, we reduced the GPT-4o input to 30 poems per prompt, resulting in a successful analysis in 17 out of 20 attempts. Further reduction to 15 poems per prompt increased the success rate to 100%. Claude-3-Opus also successfully ranked the 15-poem prompts on every attempt.

Therefore, we will present the evaluations of 90 poems in one prompt for Claude-3-Opus only, and evaluations relying on 15 poems in one prompt for both Claude-3-Opus and GPT-4o.

Model	Metric	Innovative	Poeticness	Creativity	Quality	Similarity
Claude-3 90n	SRC	0.63	0.67	0.63	0.57	0.47
	p-val	2.1e-11	7.36e-13	2.1e-11	5.8e-09	3.5e-06
Claude-3 90	SRC	0.68	0.63	0.62	0.57	0.52
	p-val	1.2e-13	2.1e-11	9.7e-11	5.8e-09	1.9e-07
Claude-3 15n	SRC	0.7	0.8	0.72	0.87	0.83
	p-val	1.6e-14	3.2e-21	2e-15	2.6e-28	2.2e-24
Claude-3 15	SRC	0.67	0.78	0.68	0.85	0.8
	p-val	7.4e-13	7.2e-20	1.2e-13	3.2e-26	3.2e-21
GPT-4o 15n	SRC	0.75	0.68	0.72	0.82	0.7
	p-val	1.8e-17	1.2e-13	2e-15	1e-22	1.6e-14
GPT-4o 15	SRC	0.73	0.7	0.77	0.77	0.72
	p-val	2e-16	1.6e-14	1.3e-18	1.3e-18	2e-15

Table 17: Results of Spearman’s Rank Correlation (SRC) against the ground truth of the poems ordering in 5 criteria from 90-poems prompts using Claude-3-Opus (Experiment 2, “Claude-3-Opus 90n” and “Claude-3-Opus 90”) and 15-poems prompts (Experiment 3, “Claude-3-Opus 15”, “Claude-3-Opus”, “GPT-4o 15n” and “GPT-4o”). The best result for each model is presented in bold. For comparison with human evaluations please see Table 14.

6.5.2 Experiment 2—Evaluation of Poems—90 poems in a prompt

Since the GPT-4o version used in this study was incapable of evaluating all 90 poems in a single prompt, this experiment is conducted only with Claude-3-Opus. For prompts, we have prepared 10 datasets containing all 90 poems, where the poems' ordering for each dataset was randomized. Thus, each poem will be evaluated ten times.

Each of the ten 90-poem datasets underwent evaluation in each of the specified five criteria: Creativity, Quality, Innovativeness, Similarity, and Poeticness, using the prompts presented in Section 6.1. Post-evaluation, the numerical scores for each poem were averaged, and the poems were ranked based on their average scores. We have used both the original 1–5 scale scores assigned by the model (Claude-3-Opus 90), and the poem's positions on the list (Claude-3-Opus 90n).

After sorting the poems by their average evaluation score, each poem's position was marked with the category it represents (A, B or C), thus producing an ordered list of A's, B's and C's. These lists were then compared to the ground truth (30 *times* A, followed by 30 *times* B, followed by 30 *times* C), and Spearman's rank correlation was computed, using the ranks A, B, and C. Table 17 presents Spearman's correlation coefficients calculated against the ground truth order, and Figure 18 presents graphical representations of the results. These outcomes demonstrate an even higher degree of correlation with the ground truth (SRC=0.68) as compared to "without-context" evaluations, and are all statistically significant, with very low p-values in every case. However, we note that some of the 90-poem evaluations failed and had to be repeated. Subsequent experiments will show that reducing the number of poems in the prompt improves the accuracy even further.

6.5.3 Experiment 3—Evaluation of Poems—15 poems in a prompt

In this experiment, 5 poems were randomly selected from each category ("Good"=A, "Medium"=B and "Bad"=C) to form a subset of 15 poems for each query to both LLMs. This process was repeated to create 100 unique subsets, each containing poems from all three categories and shuffled accordingly.

Each of the 100 15-poem datasets underwent evaluation using the same five criteria: Creativity, Quality, Innovativeness, Similarity, and Poeticness, using the prompts shown in Section 6.1.

Due to random sampling, the frequency of poems across the 100 subsets was slightly different, but theoretically, every poem had the same expected number of use. After evaluation, scores for each poem were averaged across all subsets where they appeared, and the poems were ranked by their average evaluation scores. This was done for both types of scoring methods.

As before, each poem's position in the scoring list was marked with its category (A, B or C), resulting in an ordered list. These lists were compared to the ground truth using Spearman's rank correlation with

Model		SRC
QUALITY ORDER		
Claude-3 15n:		0.87
Claude-3:		0.85
GPT-4o 15n:		0.82
GPT-4o:		0.77
Claude-3 90n:		0.57
Claude-3 90:		0.57
INNOVATIVENESS ORDER		
Claude-3 15n:		0.70
Claude-3 15:		0.67
GPT-4o 15n:		0.75
GPT-4o:		0.73
Claude-3 90n:		0.63
Claude-3 90:		0.68
CREATIVITY ORDER		
Claude-3 15n:		0.72
Claude-3:		0.68
GPT-4o 15n:		0.72
GPT-4o:		0.77
Claude-3 90n:		0.63
Claude-3 90:		0.62
POETICNESS ORDER		
Claude-3 15n:		0.80
Claude-3:		0.78
GPT-4o 15n:		0.68
GPT-4o:		0.7
Claude-3 90n:		0.67
Claude-3 90:		0.63
SIMILARITY ORDER		
Claude-3 15n:		0.83
Claude-3:		0.80
GPT-4o 15n:		0.70
GPT-4o:		0.72
Claude-3 90n:		0.47
Claude-3 90:		0.52
HUMAN NON-EXPERT JUDGES ORDERINGS		
NOVELTY:		0.38
IMAGIN.:		-0.12
VALUE:		-0.20
QUALITY:		-0.33
APPREC.:		-0.33
SKILL:		-0.35
TYPICALITY:		-0.37
WELLBEING:		-0.45
EFFORT:		-0.45

= Good (A) = Medium (B) = Bad (C)

Figure 18: The poem orders obtained in Experiments 2 and 3 on all criteria, where every poem is colour coded with its ground truth category. The lower part of the table presents for comparison the orderings by human non-expert judges from Lamb, Brown and Clarke (2015).

ranks A, B and C. Table 17 presents those correlations, showing a very high degree of consistency with the ground truth, and all values are statistically significant with extremely low p-values.

Overall, most results surpass those obtained by Claude-3-Opus when evaluating all 90 poems in a single prompt and significantly surpass those of human non-expert judges. Evaluations of 15 poems per prompt yielded much higher correlations to the ground truth. The highest correlation, 0.87, was achieved in the Quality criterion by Claude-3-Opus 15n (Table 17 and Figure 18) when the scores were the poem's positions in the list. For GPT-4o, the results were slightly lower, but the Quality criterion also yielded the highest correlation when the scores were the poem's positions in the list. Deriving the scores from the poem's positions consistently improved results for Claude-3-Opus, but this was not always the case for GPT-4o (Table 17 and Figure 18).

Analysis of Variance

Similar to the approach taken by Lamb, Brown and Clarke (2015), we conducted a single-factor ANOVA for each evaluation criterion to compare the average scores of the “Good,” “Medium,” and “Bad” poem categories, as shown in Table 18 for all our “in-context” evaluations. Given that we had five evaluation criteria, we performed five separate ANOVAs for each model. The null hypothesis stated that all poem categories have the same population mean, while the alternative hypothesis suggested that significant differences exist among these means.

The results, presented in Table 18, indicate statistically significant differences between the poem category means for each criterion, considering the Bonferroni-corrected alpha level of 0.01. This finding implies that there are significant variations among the groups for all tested criteria, suggesting that a poem's category has a measurable impact on the scores of all five criteria.

The “Bad” category plays an important role in validating the sensitivity of the LLM evaluators. As shown in Table 18, the drop-off in mean scores from the “Medium” to the “Bad” category is precipitous across all models and criteria (e.g., for Claude-3-Opus 15n Quality, the score drops from 9.06 to 3.33). This sharp distinction confirms that the models are not merely reacting to surface-level features like formatting or length, but are successfully detecting the fundamental deficiencies in craft (such as forced rhymes, cliché, or lack of coherence) that characterized the “Bad” dataset. The ability of the LLMs to decisively separate this bottom tier from the professional tiers confirms that their scoring reflects a broad spectrum of poetic quality. This demonstrates that the models are not merely distinguishing between high-level stylistic nuances found in professional works, but are capable of correctly identifying and penalizing fundamental deficiencies in amateur writing.

We can also observe that using the poem's position in the output list as a scoring method increases the F-value and decreases the p-value in every case, indicating higher effect size of this approach.

Criterion	Good	Medium	Bad	F-value	p-value
GPT-4o					
Average Innovativeness Score	3.49	3.12	1.64	82.72	7.5e-21
Average Quality Score	3.48	2.98	1.71	92.31	3.1e-22
Average Creativity Score	3.46	3.15	1.77	78.8	2.96e-20
Average Poeticness Score	3.82	3.42	2.15	60.27	3.75e-17
Average Similarity Score	3.41	2.86	2.13	49.28	4.9e-15
GPT-4o 15n					
Average Innovativeness Score	10.99	9.57	3.63	108.06	2.62e-24
Average Quality Score	11.32	9.30	3.50	157.45	1.23e-29
Average Creativity Score	10.81	9.63	3.79	96.72	7.73e-23
Average Poeticness Score	10.72	9.11	4.14	71.43	4.42e-19
Average Similarity Score	10.35	8.29	5.28	57.56	1.19e-16
Claude-3-Opus					
Average Innovativeness Score	3.24	2.77	1.54	36.3	3.44e-12
Average Quality Score	3.52	2.84	1.43	94.47	1.56e-22
Average Creativity Score	3.49	2.93	1.64	50.04	3.43e-15
Average Poeticness Score	3.85	3.2	1.8	72.02	3.53e-19
Average Similarity Score	3.37	2.69	1.58	90.5	5.55e-22
Claude-3-Opus 15n					
Average Innovativeness Score	10.83	9.02	4.26	52.89	9.3e-16
Average Quality Score	11.65	9.06	3.33	132.23	4.2e-27
Average Creativity Score	11.11	8.95	4.02	62.55	1.46e-17
Average Poeticness Score	11.42	8.81	3.79	85.15	3.27e-21
Average Similarity Score	11.22	8.63	4.13	111.08	1.11e-24
Claude-3-Opus 90n					
Average Innovativeness Score	56.5	48.9	31.1	41.2	2.64e-13
Average Quality Score	53.7	50.9	32.9	25.74	1.66e-9
Average Creativity Score	54.6	48.1	33.8	31.22	6.02e-11
Average Poeticness Score	56.2	48.6	31.7	35.9	4.29e-12
Average Similarity Score	51.2	47.8	37.5	16.13	1.1e-6
Claude-3-Opus 90					
Average Innovativeness Score	3.04	2.82	2.26	30.09	1.96e-12
Average Quality Score	2.6	2.55	1.92	24.57	3.47e-9
Average Creativity Score	3.13	2.91	2.51	19.84	7.94e-8
Average Poeticness Score	2.94	2.67	1.97	37.34	1.96e-12
Average Similarity Score	2.56	2.41	2.11	14.95	2.63e-6

Table 18: ANOVA results for poem evaluation in five criteria in Experiments 2 and 3 for all models with both scoring methods. Columns “Good,” “Medium,” and “Bad” present averaged scores for those poems’ categories.

6.6 Experiment 4—Evaluating Poetry - Interrater Reliability of Claude-3-Opus and GPT-4o

The primary goal of the previous experiments was to approximate the ground truth categorization. Now, we will test if the LLMs return different rankings when the same set-up is repeated several times. For this experiment, the LLM’s temperature parameter was set to 1 to maximize diversity, as with all other experiments in this chapter. After that, we check if the diverse outputs are consistent and reliable with respect to the rankings they yield. For that we conducted a dedicated experiment where we have randomly selected 1 subset of 15 poems (from the subsets used in Experiment 3) and evaluated this subset 10 times with our LLMs. We are evaluating both models with both scoring methods. Effectively, 200 evaluations were conducted (2 models \times 2 ways of scoring \times 10 repetitions \times 5 criteria).

In Table 19 we present an example of evaluating the “Creativity” criterion with Claude-3-Opus 15n and Claude-3-Opus, where the outputs are different, yet similar, across multiple runs of the same set-up.

Knowing that the outputs are diverse, we gauged the reliability of these LLM-based evaluations, and explored a range of Intraclass Correlation Coefficient (ICC) algorithms (Shrout and Fleiss 1979). We considered ICC models that cater to single measurements with fixed rater effects (ICC1), models treating raters as random effects from a larger pool (ICC2), and those considering systematic differences between raters (ICC3), along with their k-rater variants for average ratings. Given the experimental design leveraging repeated LLM queries, and reliance on the averaged scores, the “k” versions of the tests seem the most appropriate reliability measures, which we present in Table 20.

The correlation results of ICC1k, ICC2k and ICC3k were all very similar, and they ranged between 0.9 and 0.99 for evaluations of in all 5 criteria. In every case, using the scores derived from the poem’s position in the output list has significantly lowered the p-values and increased the F-scores. All the tests show very high correlation between raters and all results are statistically significant (very low p-values).

Overall, these results confirm that while the temperature parameter successfully diversifies the output of LLMs (Table 19), the ICC results show that those diversified outputs are stable and reliable on average (Table 20).

6.7 Comparison with Non-Expert Human Evaluations

The progression of LLM performance in poetry evaluation demonstrates a clear superiority over non-expert human judges across multiple experiments.

In the “without-context” evaluation (Experiment 1), both Claude-3-Opus and GPT-4o already surpassed human performance, achieving correlations of 0.57 and 0.62 respectively. This improvement was

achieved despite the models' tendency toward uneven distribution of categories, particularly their bias toward middle-category assignments. The progression of GPT-4 versions showed steady improvement, with the latest GPT-4o demonstrating the strongest correlation with ground truth.

The transition to “in-context” evaluations yielded even more substantial improvements. When evaluating 90 poems simultaneously, Claude-3-Opus achieved correlations ranging from 0.47 to 0.67, depending on the criterion and scoring method. These results consistently outperformed the human baseline across all evaluation criteria. The most significant improvements emerged in Experiment 3, where both models evaluated 15-poem subsets. Claude-3-Opus achieved its highest correlation of 0.87 for the “Quality” criterion, while GPT-4o reached 0.82, both substantially exceeding human performance.

The high reliability of LLM evaluations is further evidenced by the very high interrater reliability results from Experiment 4. Both Claude-3-Opus and GPT-4o demonstrated high ICC values ranging between 0.9 and 0.99 across all five evaluation criteria, with statistically significant p-values. These ICC scores, which exceed the typical acceptable threshold of 0.7-0.9 for human judges in CAT studies, indicate that despite the temperature parameter introducing variability in the models' outputs, their evaluations remain highly consistent and reliable across multiple iterations (Table 20). This level of reliability suggests that LLMs can provide stable and reproducible poetry assessments, addressing one of the key challenges in creative evaluation - the need for consistent judgment across multiple evaluations.

These findings demonstrate that LLMs have already achieved remarkable capabilities in poetry evaluation with a meaningful understanding of poetic quality, consistently surpassing non-expert human judges across multiple criteria and experimental conditions. The combination of strong ground truth correlation and high inter-rater reliability across different evaluation methods validates their potential for reliable creative assessment.

6.8 Further Discussion and Implications

The success of Claude-3-Opus and GPT-4o in evaluating poetry against ground truth benchmarks, consistently surpassing non-expert human judges as demonstrated in this chapter, suggests that the CAT-inspired, in-context LLM evaluation methodology presented here is a promising approach likely to maintain relevance. As LLM capabilities continue their rapid evolution, this framework offers a foundation for future automated creative assessment.

Anticipated advancements in next-generation LLMs are poised to further enhance the utility of this methodology. For instance, an improved understanding of complex social, cultural, and historical contexts could enable future models to better account for factors currently challenging for automated systems. A critical question for future investigation is whether such advancements will allow LLMs,

when applying this framework, to eventually match or even exceed the nuanced evaluations of domain experts, moving beyond the non-expert benchmark already surpassed in here. Achieving such parity would pose profound questions about the role of computational systems in formal artistic assessment and curation. Crucially, these anticipated model improvements would likely bolster the effectiveness of our in-context, multi-criteria evaluation framework rather than rendering it obsolete; its core principles of relative assessment and forced comparison address fundamental measurement challenges independent of specific model capabilities.

Furthermore, as LLMs become more influential evaluators, ensuring aesthetic diversity becomes paramount. The observed consensus between Claude-3-Opus and GPT-4o could reflect shared training data biases or architectural similarities. Future work must focus on incorporating diverse aesthetic traditions and cultural perspectives into LLM evaluation frameworks to avoid inadvertently reinforcing narrow or homogenized notions of literary quality. Developing methods to probe and mitigate these biases will be crucial for the responsible application of LLMs as evaluators in creative domains.

6.9 Conclusion

Our experiments highlight the significant potential of state-of-the-art Large Language Models, specifically Claude-3-Opus and GPT-4o, to serve as effective evaluators of poetry. We demonstrated that these models, when utilized within our CAT-inspired in-context framework, achieve assessments that show strong correlations with ground truth quality benchmarks derived from publication venues, markedly outperforming non-expert human judges in this regard. Furthermore, the high inter-rater reliability observed across multiple criteria underscores the consistency and robustness of this LLM-based evaluation approach.

However, while the results demonstrate high correlation with the ground truth, several critical limitations regarding the nature of the classification and the evaluators must be addressed. First, the concept of a ground truth in poetry is inherently problematic. The tripartite classification (“Good”, “Medium”, “Bad”) assumes a hierarchy that may not exist in reality; poetry often functions on multiple dimensions—emotional resonance, technical craft, and semantic density—which do not always align on a single linear scale. A poem might be technically flawed yet emotionally profound, or technically perfect yet creatively inert. By collapsing these dimensions into a single score, both the human ground truth and the LLM evaluations inevitably lose nuance.

Second, the comparison with non-expert human judges, while methodologically necessary due to the available dataset, sets a specific baseline. Surpassing non-experts confirms that LLMs have acquired a sophisticated representation of poetic norms and can identify technical competence. However, this

does not guarantee that LLMs would perform equivalently against expert human critics, who might value transgressive or avant-garde qualities that statistical models, trained to favor probability and coherence, might penalize.

Finally, LLMs possess significant inherent restrictions as evaluators of creative work. They lack lived experience, sensory perception (qualia), and genuine emotional capacity. When an LLM evaluates a poem about grief, it recognizes patterns of language associated with grief in its training data, but it cannot resonate with the human experience of loss. This raises the question of whether they are evaluating “poetry” as an expressive act, or merely “poeticness” as a linguistic texture. The high reliability scores observed in our experiments indicate consistency in detecting these linguistic textures, but they should not be conflated with the human capacity for artistic appreciation.

Finally, we must address the potential for data contamination, particularly regarding the poems in the “Good” category. As these works were sourced from a prominent publication (*Poetry* magazine), it is highly probable that they exist within the pre-training corpora of models as large as Claude-3-Opus and GPT-4o. Consequently, high quality scores assigned to these poems could stem, in part, from the models recognizing these texts as “published” or “famous” rather than solely from an intrinsic analysis of their quality. However, we deliberately chose to retain the exact dataset used by Lamb, Brown and Clarke (2015) rather than introducing new, contamination-free poems. This decision was driven by the necessity of the comparative benchmark; introducing new poems would have invalidated the direct comparison with the non-expert human evaluation data recorded in the 2015 study. Thus, the finding that LLMs align closer to ground truth than non-experts remains valid within the context of reproducing and auditing that specific experimental setup.

Subject to these limitations, this study offers the following technical insights into the capabilities of LLMs for automated creative assessment:

1. We introduced a **novel LLM-based evaluation methodology** inspired by the Consensual Assessment Technique (CAT), finding that it can provide consistent assessments that align with publication-based quality tiers, offering a scalable approximation of consensus judgment.
2. Our experiments suggest that evaluating **smaller, curated subsets** of poems tends to yield more reliable differentiation between quality tiers compared to assessing large datasets in a single prompt.
3. We found evidence that using a **rank-based scoring methodology** (deriving scores from a poem’s position in the evaluator’s output list) can enhance alignment with established benchmarks compared to relying solely on raw numerical scores.

Having established and validated this LLM-based framework as a reliable tool for assessing poetic quality in this chapter, the crucial next step is to apply it to the core objective of evaluating AI-generated poetry

itself. This validated methodology provides the necessary foundation to directly compare the outputs of contemporary AI poetry generation techniques against established human benchmarks. Therefore, Chapter 7 employs the evaluation methods developed and confirmed here to conduct a systematic comparison, assessing the quality of poems generated by state-of-the-art models (using both prompting and fine-tuning) relative to human-written poems from different quality tiers. This allows us to leverage the insights gained about reliable LLM evaluation to directly address questions about the current capabilities of AI in creative poetry generation.

While future work should continue exploring scalability, refining criteria, and comparing against expert human judgment, this chapter provides a robust methodology and compelling evidence for the utility of LLMs in creative assessment. It lays the groundwork for applying these techniques, as done in the subsequent chapter, and potentially extending them to practical applications in literary evaluation and publishing.

Claude-3-Opus 15n												
Poem ID	Cat.	Scores										Average Score
Poem 27	A	14	15	15	14	15	15	14	15	15	14	14.6
Poem 3	A	15	13	12	13	14	10	15	11	13	11	12.7
Poem 7	A	13	14	9	11	13	14	13	14	14	12	12.7
Poem 6	A	12	12	13	15	12	11	8	12	11	15	12.1
Poem 8	A	11	9	14	10	10	13	7	13	12	8	10.7
Poem 50	B	9	7	10	12	9	12	11	8	8	9	9.5
Poem 54	B	7	11	7	6	11	9	12	9	9	13	9.4
Poem 53	B	10	10	11	9	8	8	10	7	10	10	9.3
Poem 41	B	8	8	8	8	7	7	6	10	7	6	7.5
Poem 42	B	6	6	5	7	6	6	9	5	5	5	6.0
Poem 61	C	5	5	6	5	5	5	5	6	6	7	5.5
Poem 74	C	4	3	3	4	4	2	4	4	4	4	3.6
Poem 79	C	3	2	2	3	3	1	3	2	3	1	2.3
Poem 65	C	2	1	4	2	2	3	1	3	2	3	2.3
Poem 69	C	1	4	1	1	1	4	2	1	1	2	1.8

Claude-3-Opus 15												
Poem ID	Cat.	Scores										Average Score
Poem 27	A	4	5	5	4	5	5	4	5	5	4	4.6
Poem 3	A	5	3	4	4	4	3	5	3	4	3	3.8
Poem 7	A	4	4	3	3	4	4	4	4	4	3	3.7
Poem 6	A	3	2	4	5	3	3	3	3	3	5	3.4
Poem 50	B	3	1	4	4	3	4	3	3	3	3	3.1
Poem 53	B	3	1	4	3	3	3	3	3	3	3	2.9
Poem 54	B	3	1	3	2	3	3	3	3	3	4	2.8
Poem 8	A	3	1	4	3	3	4	2	3	3	2	2.8
Poem 41	B	3	1	3	3	3	2	2	3	2	2	2.4
Poem 42	B	2	1	2	2	2	2	3	3	2	2	2.1
Poem 61	C	2	1	3	2	2	2	2	3	2	2	2.1
Poem 74	C	2	1	2	2	2	2	2	2	2	2	1.9
Poem 65	C	2	1	2	1	2	2	1	2	1	1	1.5
Poem 79	C	2	1	2	1	2	1	1	2	1	1	1.4
Poem 69	C	1	1	1	1	1	2	1	1	1	1	1.1

Table 19: Example of 10 evaluations of the same subset of 15 poems in Experiment 4 with Claude-3-Opus 15n and Claude-3-Opus, evaluating the criterion “Creativity”. Poem’s numbers indicate the categories they belong to: 1-30 “Good” (A), 31-60 “Medium” (B), 61-90 “Bad” (C). These results confirm that while the temperature parameter successfully diversifies the output of its LMM, the ICC results Table. 20 show that those diversified outputs are consistent and reliable on average.

Criteria & Model	ICC1k			ICC2k			ICC3k		
	ICC	F	p-value	ICC	F	p-value	ICC	F	p-value
CREATIVITY									
Claude-3-Opus	0.98	54.66	1.63e-48	0.98	53.59	4.14e-46	0.98	53.59	4.14e-46
Claude-3-Opus 15n	0.99	80.84	2.12e-58	0.99	75.45	3.4e-54	0.99	75.45	3.4e-54
GPT-4o	0.97	36.25	8.17e-39	0.97	42.59	6.54e-41	0.98	42.59	6.54e-41
GPT-4o 15n	0.99	92.32	7.22e-62	0.99	86.17	1.98e-57	0.99	86.17	1.98e-57
INNOVATIVE									
Claude-3-Opus	0.99	101.03	2.99e-64	0.99	105.98	1.47e-62	0.99	105.98	1.47e-62
Claude-3-Opus 15n	0.99	161.46	6.16e-77	0.99	150.7	1.74e-71	0.99	150.7	1.74e-71
GPT-4o	0.98	44.61	1.28e-43	0.98	51.25	4.41e-45	0.98	51.25	4.41e-45
GPT-4o 15n	0.99	75.05	1.75e-56	0.99	70.05	2.09e-52	0.99	70.05	2.09e-52
POETICNESS									
Claude-3-Opus	0.99	98.95	1.07e-63	0.99	102.67	9.13e-62	0.99	102.67	9.13e-62
Claude-3-Opus 15n	0.99	147.33	2e-74	0.99	137.51	3.83e-69	0.99	137.51	3.83e-69
GPT-4o	0.98	43.34	6.17e-43	0.98	49.62	2.4e-44	0.98	49.62	2.4e-44
GPT-4o 15n	0.99	94.85	1.4e-62	0.99	88.52	4.31e-58	0.99	88.52	4.31e-58
QUALITY									
Claude-3-Opus	0.99	93.41	3.56e-62	0.99	98.77	8.39e-61	0.99	98.77	8.39e-61
Claude-3-Opus 15n	0.99	156.2	5e-76	0.99	145.79	1.23e-70	0.99	145.79	1.23e-70
GPT-4o	0.95	21.74	5.24e-28	0.95	29.67	3.28e-33	0.97	29.67	3.28e-33
GPT-4o 15n	0.99	83.52	3e-59	0.99	77.96	5.5e-55	0.99	77.96	5.5e-55
SIMILARITY									
Claude-3-Opus	0.94	15.46	9.2e-22	0.94	15.66	2.2e-21	0.94	15.66	2.2e-21
Claude-3-Opus 15n	0.96	25.74	2.1e-31	0.96	24.03	4.93e-29	0.96	24.03	4.93e-29
GPT-4o	0.9	10.37	1.46e-15	0.9	9.9	1.35e-14	0.9	9.9	1.35e-14
GPT-4o 15n	0.92	12.06	9.27e-18	0.92	11.26	2.42e-16	0.91	11.26	2.42e-16

Table 20: Results of ICC evaluations in Experiment 4, computed for both models with both scoring approaches and for the five criteria.

Chapter 7

Human Poetry vs AI Poetry: LLM Evaluation

As the capstone of this thesis, this chapter applies the LLM-based evaluation framework established in Chapter 6 to the task of assessing the perceived *quality* of poetry generated by contemporary AI. Conducted in mid-2024, this investigation leverages Claude-3-Opus and GPT-4o as evaluators to compare AI-generated poetry against established human benchmarks.

The primary objective is to compare the *quality* of AI-generated poems, produced via two distinct techniques, against the human quality tiers (“Good”, “Medium”, and “Bad”) from the dataset introduced in Chapter 6. The AI generation methods explored are:

1. **Prompt-based generation:** Utilizing the direct instruction-following capabilities of Claude-3-Opus and GPT-4o to generate poems from summaries.
2. **Fine-tuning:** Applying the summary-poem pair dataset structure (developed in Chapter 4 for stylistic control) to fine-tune GPT-4o models. Here, fine-tuning serves solely as an alternative generation method for *quality* comparison, not for specific stylistic imitation.

We evaluate both direct “raw” outputs from fine-tuned models and “refined” versions subjected to automated post-processing (Section 7.2.3). This distinction allows for a nuanced comparison between fine-tuning and prompt-based approaches, isolating the impact of minor grammatical imperfections on perceived quality.

Before presenting the results, it is crucial to define the scope and limitations of this comparative framework. As discussed in Chapter 6, the operationalization of poetic “quality” based on publication venue serves as a necessary ground truth for statistical analysis but inevitably flattens the multidimensional

nature of artistic merit into a linear hierarchy. Furthermore, the use of LLMs as evaluators introduces distinct constraints; these models lack lived experience, sensory perception (qualia), and emotional capacity. Consequently, the evaluations presented in this chapter represent a measure of “LLM-perceived quality”—a metric that assesses linguistic proficiency, coherence, and adherence to learned patterns of literary excellence, but which should not be conflated with the full depth of human artistic appreciation.

Two central findings emerge from this chapter. Firstly, poems generated via prompting alone often achieve quality ratings that match or exceed those of “Good” and “Medium” human poems, according to our LLM evaluators. This underscores the remarkable inherent capabilities of current general-purpose models for producing high-quality poetic text. Secondly, and critically, a direct comparison of evaluations performed by Claude-3-Opus and GPT-4o reveals notable discrepancies in their assessment tendencies. While both demonstrate high reliability and agree on certain aspects (e.g., consistently ranking “Bad” human poems lowest), Claude-3-Opus exhibits a markedly stronger preference for AI-generated poems than GPT-4o, often rating them significantly higher than even “Good” human poems. This highlights the crucial importance of considering potential evaluator biases when interpreting LLM-based assessments of creative work.

Specifically, this chapter contributes:

- Application of the validated LLM-based evaluation framework (Chapter 6) to quantify the perceived merit of AI-generated poetry relative to tiered human benchmarks.
- A direct comparison of prompt-based versus fine-tuning generation methods (evaluating both raw and refined outputs) using GPT-4o, assessing their effectiveness for producing high-caliber poetic output, irrespective of specific style imitation.
- A comparative analysis of evaluation patterns between Claude-3-Opus and GPT-4o, identifying consensus points and significant differences in their assessment biases towards AI-generated content.

The chapter proceeds by detailing the datasets (Section 7.1) and generation methods (Section 7.2), including the selection process for evaluation based on initial quality screening (Section 7.3). Subsequent sections present the core comparative quality evaluations: AI versus human poetry (Section 7.4), comparisons across different AI generation methods (Section 7.5.1), including analysis of top versus bottom quality performers (Section 7.5.3), and inter-rater reliability assessment for quality judgments (Section 7.6). Finally, Section 7.7 synthesizes the findings, discusses implications including the observed evaluation discrepancies, and concludes the chapter.

7.1 Datasets

7.1.1 Fine-tuning dataset

Author	Birth	Death	Nationality	Poems in Fine-tuning Set	Poems in Test Summary Set
Ella Wheeler Wilcox	1850	1919	American	511	22
William Butler Yeats	1865	1939	Irish	324	12
Thomas Hardy	1840	1928	English	270	20
Amy Lowell	1874	1925	American	166	7
Katharine Tynan	1859	1931	Irish	146	7
Ivor Gurney	1890	1937	English	106	4
Edward Thomas	1878	1917	Welsh/English	86	6
Rupert Brooke	1887	1915	English	75	6
Wilfred Owen	1893	1918	English	65	2
Isaac Rosenberg	1890	1918	English	57	3
Francis Ledwidge	1887	1917	Irish	44	2
Edith Wharton	1862	1937	American	32	2
Alan Seeger	1888	1916	American	30	1
Gertrude Stein	1874	1946	American	29	4
John McCrae	1872	1918	Canadian	27	1
Laurence Binyon	1869	1943	English	23	1
Charles Sorley	1895	1915	Scottish	9	0

Table 21: Demographic information and poem counts by author for fine-tuning and test summary datasets. The “Poems in Fine-tuning Set” column indicates the number of poems by each author included in the 2000-poem fine-tuning corpus. The “Poems in Test Summary Set” column indicates the number of poems by each author among the 100 poems reserved for generating test summaries.

To fine-tune the GPT-4o models, we curated a dataset of 2,100 poems from acclaimed English-language poets, primarily from the 19th and early 20th centuries (Table 21). Sourced from Project Gutenberg (Project Gutenberg 2024) and the Poetry Foundation (Poetry Foundation 2024), authors were selected based on literary prominence, public domain status (deceased >75 years), and availability of work. To manage processing demands, poem length was capped at 500 words.

The authors whose works constitute this fine-tuning dataset are detailed in Table 21, which provides their birth and death years, and nationality. This selection encompasses a range of influential voices from the period. Utilizing a broader corpus from multiple authors for fine-tuning in this chapter aimed to equip the models with a diverse range of high-quality poetic patterns, enabling them to generate novel poems based on summaries for subsequent quality assessment, rather than focusing on mimicking a single

author’s specific style. Each poem from these authors was summarized using Claude-3-Opus to create a corresponding synopsis intended to capture its essential themes and content, utilizing the following prompt:

```
You will be given a poem to summarize. Your task is to write a concise
summary of the poem, not exceeding 100 words. Here is the poem:
```

```
<poem>
```

```
[POEM]
```

```
</poem>
```

```
Please read and analyze the poem carefully. In your summary, focus
on the following elements:
```

1. The main theme or central idea of the poem
2. Key imagery or symbolism used
3. The overall mood or tone

```
Remember to capture the essence of the poem without simply paraphrasing
it. Your summary should give readers a clear understanding of what
the poem is about and how it conveys its message.
```

```
Remember, your summary must be no longer than 100 words.
```

```
Provide your summary inside <summary> tags. Do not include any other
text outside these tags.
```

From the collected poems, 2,000 were randomly selected to form the fine-tuning dataset for the GPT-4o models. The remaining 100 poems were reserved for generating summaries that will be used to create new poems, as described in the following sections.

The fine-tuning dataset comprises of 2000 “[poem summary] — [poem]” pairs following the methodology presented in Chapter 4 (Section 4.3.2).

7.1.2 Test Summaries for Poem Generation

In addition to the fine-tuning dataset, 100 summaries were produced from the reserved set of 100 poems. These summaries are part of the prompts for generating new poems using both the fine-tuned GPT-4o models and the prompt-only generation method with Claude-3-Opus and GPT-4o. By ensuring that these

test summaries were not included in the fine-tuning dataset, we aimed to evaluate the models' abilities to generate novel poems based on unseen summaries.

7.2 Methods for Generating Poems

7.2.1 Generating Poems Through Prompts Only

We employed Claude-3-Opus and GPT-4o to generate poems using the 100 test summaries described in Section 7.1.2 as inputs. To ensure consistency across generation methods and alignment with standard short poetic forms, we imposed a maximum length of 20 lines. Following generation, unique fictional author names were assigned to each poem to facilitate blind evaluation. The models were instructed using the specific prompt below:

```
You will be given a summary of a poem. Your task is to write an original poem based on this summary. Here are the specific instructions:
```

```
1. Read the following poem summary carefully:
```

```
<poem_summary>
```

```
[SUMMARY]
```

```
</poem_summary>
```

```
2. Write an original poem inspired by this summary. The poem should capture the essence, mood, and key elements described in the summary.
```

```
3. The poem should have no more than 20 lines, excluding blank lines. You may use fewer lines, but do not exceed 20 lines.
```

```
4. Generate a title for the poem that reflects its content or theme.
```

```
5. Create a filename for the poem using the following format: "Poem Title.txt". For example, if the title is "Sadness amongst the stars", the filename would be "Sadness amongst the stars.txt".
```

```
6. Format your output as follows:
```

```
- Place the poem without its title within <poem> tags.
```

```
- Place the filename within <filename> tags.
```

```
Remember to be creative and original while staying true to the essence of the provided summary. Your poem should evoke emotions and paint vivid imagery related to the themes presented in the summary.
```

7.2.2 Generating Poems with Fine-Tuned GPT-4o Models

Two models, GPT-4o-2024-08-06 and GPT-4o-mini-2024-07-18, were fine-tuned using the methodology outlined in Chapter 4. We utilized default hyperparameters over three epochs, training on structured “summary-poem” pairs. This approach conditions the model to adhere to the thematic content of a summary while implicitly learning stylistic patterns from the training corpus. These fine-tuned models were then tasked with generating poems from the 100 test summaries; generation was repeated if outputs exceeded the 20-line limit until a complete dataset was formed.

7.2.3 Refining AI-generated Poems from Fine-Tuned Models

Initial outputs from the fine-tuned models occasionally exhibit minor grammatical errors or awkward phrasing not present in the base models. To ensure the subsequent quality comparison focused on substantive poetic merit rather than surface-level imperfections, we implemented an automated refinement step. We employed the base GPT-4o model to correct grammar and flow while strictly preserving the original essence and intent of the fine-tuned output. We present results for both “raw” and “refined” versions to assess the specific impact of this post-processing.

The prompt used for this refinement process is as follows:

```
You are a skilled poetry editor tasked with improving a given poem. Your goal is to enhance the poem's language, grammar, and logical consistency while maintaining its original essence and intent. Follow these steps:
```

```
1. First, carefully read the following poem:
```

```
<poem>
[POEM]
</poem>
```

```
2. Analyze the poem, paying close attention to:
```

- The overall theme and message
- The structure (e.g., rhyme scheme, meter, stanza format)
- The use of literary devices (e.g., metaphors, similes, alliteration)

```
3. Improve the language and grammar:
```

- Correct any spelling or grammatical errors
- Enhance word choice to better convey the poem's meaning

- Ensure proper punctuation and capitalization
 - Maintain consistency in tense and point of view
4. Address any logical inconsistencies:
 - Identify any contradictions or unclear ideas within the poem
 - Clarify confusing metaphors or analogies
 - Ensure that the poem's content or argument flows logically from beginning to end
 5. Rewrite the improved version of the poem, preserving its original structure and intent as much as possible. Present your improved version within `<improved_poem>` tags.
 6. After the improved poem, provide a brief explanation of the changes you made and why. Include this explanation within `<explanation>` tags.

Remember to maintain the poem's original voice and style while making improvements. Your goal is to enhance the poem's clarity and impact without altering its fundamental character.

It is important to note that the initial selection of the top poems (detailed in Section 7.3) was performed on the “raw” outputs. Refinement was applied subsequently to these selected poems for specific comparative experiments. Tables 22 and 23 illustrate the nature of changes introduced by this process.

7.3 Selection of AI-Generated Poems for Evaluation

To ensure the comparative evaluations focused on the strongest capabilities of each method, we implemented a dedicated selection process. We generated an initial pool of 100 poems from each of the four primary sources: Claude-3-Opus (prompt-only), GPT-4o (prompt-only), and the “raw” outputs from fine-tuned GPT-4o and GPT-4o-mini. “Refined” versions of the fine-tuned sets were subsequently created via the post-processing step described in Section 7.2.3.

We applied the evaluation methodology established in Chapter 6 to identify the top 30 poems from each generative source. For each set of 100 poems, we created 120 evaluation subsets containing 15 randomly selected poems. These were evaluated on the holistic “Quality” criterion using rank-based scoring (where rank 1 receives 15 points, down to rank 15 receiving 1 point), as this method previously yielded the highest correlation with ground truth.

Poem Generation Summary

This poem explores the consequences of misguided desires and the wisdom of contentment. The speaker recounts praying for riches, glory, and love, only to find that each granted wish brings unforeseen burdens. Wealth increases cares, fame brings pain, and love leaves scars. The imagery of gold, sung names, and devouring fire symbolizes these fleeting pleasures. The tone shifts from regret to enlightenment as the speaker finally prays for a contented mind, receiving peace and strength. The poem’s message emphasizes the importance of inner satisfaction over external achievements, suggesting that true fulfillment comes from a peaceful spirit rather than worldly success.

FT-GPT-4-mini raw	FT-GPT-4-mini refined
<p>Once I prayed for gold, and higher than my head Strong walls I built against the battering years; I did not pray for bread, and was not fed. Now I pray for strength and the huge weight of cares Lies light, - as light as dew, on my strong shoulders; Strength is a golden staff whereon the strong leans.</p> <p>I prayed for glory; the world wrote my name In the vital honey of song. In the dark years My name was nursed on the lips of men, and fed To be lust and poison and a serpent flame. Now silent I pray; the name all men use Kindles no whiteness, nor dies on their lips.</p> <p>Then I prayed for love, of love the only food - Rimmed joy and crown and instinct of my days. I had it, for it also is a passionate fire: Now I pray for a mind sleek as a root, Its sap no quarrel with the flowering heavens. I keep the scar of the love I will not slay.</p>	<p>Once I prayed for gold, and higher than my head Strong walls I built against the battering years; I did not pray for bread, and went unfed. Now I pray for strength, and the huge weight of cares Lies light — as light as dew — on my broad shoulders; Strength is a golden staff on which the strong lean.</p> <p>I prayed for glory; the world inscribed my name In the vital honey of song. In the dark years My name was nursed on the lips of men, and fed To be lust and poison and a serpent’s flame. Now silent I pray; the name all men intone Kindles no fire, nor fades upon their lips.</p> <p>Then I prayed for love, of love the only food - Rimmed joy and crown and instinct of my days. I had it, for it too is a passionate fire: Now I pray for a mind as smooth as a root, Its sap in harmony with the flowering heavens. I bear the scar of the love I will not slay.</p>

Table 22: Comparison between a raw text of a poem “From Golden Walls to Rooted Prayers” generated by FT-GPT-4-mini (based on the provided summary) and its refined version by GPT-4o.

Poem Generation Summary	
<p>This poem explores the fleeting nature of time and the bittersweet awareness of life's transience. The speaker reflects on the annual return of swifts, using their brief appearance as a metaphor for the cyclical yet ephemeral aspects of existence. The swift's "black bow" against the "harvest blue" sky creates a vivid image of summer's end. The poem's tone is melancholic and contemplative, as the speaker grapples with the realization that many experiences are fleeting and may never be repeated. The contrast between the predictable return of the swifts and the uncertainty of other life events underscores the poem's theme of impermanence and the human struggle to cherish fleeting moments.</p>	
FT-GPT-4o raw	FT-GPT-4o refined
<p>Fleetingly too I had seen A swifts' black bow on the harvest blue, And heard folly under mijn breath</p> <p>Steeling niet with fear! Now what delight To see them coming over the wall Every June to be birthed again</p> <p>Reassured and number the last week Of summer by their glass shrieks And arrow shower of glittering flight.</p> <p>Two many things were never more, Yet still the swifts came darkening air, So so the eye less tickled, yet onder the hongst of blood</p> <p>Deeper delight.</p>	<p>Fleetingly, too, I had seen A swift's black bow on the harvest blue, And heard folly under my breath.</p> <p>Stealing night with fear! Now what de- light To see them coming over the wall Every June to be born again.</p> <p>Reassured, we number the last week Of summer by their glass-like shrieks And arrow shower of glittering flight.</p> <p>Too many things were never more, Yet still the swifts came, darkening air, So the eye less tickled, yet under the hunger of blood</p> <p>Deeper delight.</p>

Table 23: Comparison between a raw text of a poem "Darkening Air Deeper Delight" generated by FT-GPT-4o (based on the provided summary) and its refined version.

For practical consistency, the selection evaluation was performed by the same model family that generated the poems (e.g., Claude-3-Opus evaluated Claude-generated poems). While this introduces a potential self-preference bias, this is mitigated in the subsequent comparative experiments (Sections 7.4 and 7.5) where **both** models act as independent evaluators across all datasets.

The 30 poems with the highest average scores from each set were selected for the final comparative study. For the fine-tuned models, the “refined” datasets consist of the post-processed versions of these same top 30 raw poems. This workflow effectively operationalizes the “Filtration” mechanism described by Ventura (2016), utilizing automated metrics to select high-value outputs from a larger generative pool. The robustness of this process was confirmed via inter-rater reliability analysis (Table 30). The highest-scoring poems from each model are presented in Section 7.3.2.

7.3.1 Addressing Plagiarized Content in Selected Fine-Tuned Poems

During the selection of the top-performing poems (Section 7.3), significant instances of plagiarism were detected among the highest-ranked outputs from the fine-tuned models. For example, the FT-GPT-4o-mini model’s top-rated poem was a verbatim reproduction of Shakespeare’s Sonnet 29. Similarly, ten of the top 30 poems initially selected from the FT-GPT-4o model were identified as plagiarized works (detailed in Table 24). Notably, several of these source texts were not present in the fine-tuning dataset, indicating that the models retrieved high-quality content from their pre-training data rather than generating novel text.

The high quality scores assigned to these plagiarized works confirm the LLM evaluators’ ability to recognize literary merit but highlight their failure to distinguish between generation and retrieval without explicit instruction. This recurrence of memorization, echoing the GPT-2 findings in Chapter 3, underscores that fine-tuning can inadvertently trigger the reproduction of existing material and demonstrates the necessity of integrating plagiarism detection into generative evaluation pipelines.

To ensure the validity of the subsequent comparative evaluations, all identified plagiarized poems were removed from the selection. The remaining non-plagiarized poems were shifted up in rank, and vacancies were filled by the next highest-scoring candidates from the original output pool. Each replacement poem underwent a manual verification process using internet search engines to ensure novelty. The analyses presented in the remainder of this chapter are based on these corrected datasets.

7.3.2 Best poems from each of the models

This section presents the highest scoring poems from each of the models, as selected by the methodology described in Section 7.3. Each poem is accompanied by the summary that was used as a prompt to generate it and the details of the original poem from which the summary was generated.

Generated Title - Initial Rank	Plagiarized Source (Author – Title/Work)	Source in Dataset?
From FT-GPT-4o-mini		
Love’s Dawn: From Despair to Hymns (1)	William Shakespeare – Sonnet 29	No
From FT-GPT-4o		
Beyond the Shining Days (6)	W. B. Yeats – The Lover Pleads with His Friend for Old Friends	Yes
Fergus and the Wandering Stars (9)	W. B. Yeats – Who Goes with Fergus?	Yes
The Poison King’s Deadly Feast (10)	A. E. Housman – Mithridates (fragment)	No
Echoes of Your Touch (13)	Conrad Aiken – Bread and Music	No
A Thousand Eyes: One Light (14)	Francis William Bourdillon – Light	No
Echoes of Time: A Moorway Reflection (21)	Thomas Hardy – Your Last Drive (first stanza)	Yes
Eternal Playmate of the Soul (24)	Mary Lamb – A Child	No
Echoes of Achievement (25)	W. B. Yeats – What Then?	Yes
Divine Grapple: Jacob’s Unexpected Victory (29)	Emily Dickinson – A little East of Jordan	No
Defying Time’s Transfiguration (30)	W. B. Yeats – The Lamentation of the Old Pensioner	Yes

Table 24: Plagiarized poems identified in the initial top selection from fine-tuned models

Claude-3-Opus: The Mountain Endures**Summary:**

This sonnet reflects on the enduring nature of a mountain and human achievement. It contrasts a recent climbing feat, where seven men conquered the peak at great cost, with the mountain’s ancient history. The poet imagines the mountain as a silent witness to cosmic and historical events, from Joshua’s biblical miracle to Caesar’s fall and Christ’s crucifixion. The imagery shifts from human scale to celestial and historical grandeur, creating a sense of awe and perspective. The tone is contemplative and reverent, emphasizing the mountain’s timeless presence against the fleeting nature of human accomplishments and historical moments.

Summary of: *Zermat: To the Matterhorn (June-July, 1897)* by Thomas Hardy.

Poem:

Silent sentinel, your stony face has seen
The rise and fall of kings and gods and men.
When Joshua’s horns rang out, you stood unmoved,
And watched impassively when Caesar bled.
From your majestic peak, the cosmos wheeled -
Millennia marked by slow, imperceptible change.
What is the span of mortal life to you?
Our greatest deeds and tragedies just wisps
Of cloud that briefly veil your timeworn crags
Before the ceaseless winds disperse them all.

Seven valiant souls now plant their flag
 Upon your crown, and stake their claim to fame.
 But you remain, as ever, granite-hewn -
 A stoic sentry at the gates of time,
 Bearing witness to the folly and the dreams
 That define the human enterprise.

GPT-4o: Hope's Enduring Flight

Summary:

This poem explores the elusive nature of Hope, personified as a fleeting entity. It suggests that Hope is not found in obvious places of joy or beauty, but rather in moments of despair and desolation. The imagery shifts from springtime optimism to stark, barren landscapes, symbolizing life's hardships. Hope is portrayed as a persistent force that endures even in the darkest times, ultimately nesting in the human heart. The tone is bittersweet, acknowledging life's struggles while affirming Hope's resilience. The poem concludes with Hope's cyclical nature, suggesting it persists across generations, finding new life in the face of death.

Summary of: A Hunting Song by Edith Wharton.

Poem:

Hope flutters not in the bloom of spring,
 Nor among laughter's easy grace;
 It dances in shadows, silent and dim,
 Nestling undaunted in the soul's vast space.
 Amidst ruins where echoes of sorrow ring,
 Where barren winds speak of lost embrace,
 Hope rises, a quiet yet fearless wing,
 Burning softly in starlit trace.
 Through deserts bleak and dreams undone,
 It weaves through dusk on threads of light,
 A whispered promise, a timeless run,
 Transforming despair into quiet might.
 Though life journeys through cycles of birth,
 In death's final sigh, the cycle renews;
 Hope finds its perch, breaching earth,

Enduring through the heart's ancient hues.

FT-GPT-4o-mini: From Golden Walls to Rooted Prayers

Summary:

This poem explores the consequences of misguided desires and the wisdom of contentment. The speaker recounts praying for riches, glory, and love, only to find that each granted wish brings unforeseen burdens. Wealth increases cares, fame brings pain, and love leaves scars. The imagery of gold, sung names, and devouring fire symbolizes these fleeting pleasures. The tone shifts from regret to enlightenment as the speaker finally prays for a contented mind, receiving peace and strength. The poem's message emphasizes the importance of inner satisfaction over external achievements, suggesting that true fulfillment comes from a peaceful spirit rather than worldly success.

Summary of: Answered Prayers by Ella Wheeler Wilcox.

Poem:

Once I prayed for gold, and higher than my head
 Strong walls I built against the battering years;
 I did not pray for bread, and went unfed.
 Now I pray for strength, and the huge weight of cares
 Lies light - as light as dew - on my broad shoulders;
 Strength is a golden staff on which the strong lean.
 I prayed for glory; the world inscribed my name
 In the vital honey of song. In the dark years
 My name was nursed on the lips of men, and fed
 To be lust and poison and a serpent's flame.
 Now silent I pray; the name all men intone
 Kindles no fire, nor fades upon their lips.
 Then I prayed for love, of love the only food -
 Rimmed joy and crown and instinct of my days.
 I had it, for it too is a passionate fire:
 Now I pray for a mind as smooth as a root,
 Its sap in harmony with the flowering heavens.
 I bear the scar of the love I will not slay.

FT-GPT-4o: Luminous Sacrament: Love's Eternal Dawn

Summary:

This sonnet explores the speaker's evolving relationship with grief. Initially feared, grief becomes a constant companion, transforming from a source of dread to a comforting presence. The poem portrays grief as a loyal "helpmeet" that accompanies the speaker through life's joys and sorrows. Imagery of travel and foreign lands symbolizes life's journey, with grief serving as an "interpreter" in difficult times. The tone shifts from apprehension to acceptance and even gratitude, as grief helps the speaker find value in challenging experiences. Ultimately, the poem presents grief as a complex, multifaceted emotion that can bring unexpected wisdom and richness to life.

Summary of: *The One Grief* by Edith Wharton.

Poem:

When first I gazed upon thy form divine,
I thought some woman's spell had lulled my sight,
And that my slumbering eyes had lost the light
That comes with dawn. Then, like the moon's design
On misty shores, thy presence did align
With infinity, a luminous might
That drew me in with hesitance so slight,
Thy gradual radiance made our path entwine.

No way but at thy will; for sacred, fine
Was thy Sacrament of Love undying,
Of all that lived our cup. Drink then, and rest,
That thy enchantment might endure and shine.
No finer blessing could a heart enshrine
To courage like the dawn's: our quest's bequest.

7.4 Comparing AI-generated Poems to Human Poetry

Having selected the top-performing AI poems from each generation method (Section 7.3), this section presents the core comparison of these AI-generated works against human-written poetry. We utilize the benchmark dataset and quality tiers (Human "Good" - A, Human "Medium" - B, Human "Bad" - C) established in Section 6.2 (Chapter 6), providing a ground truth for evaluating the AI poems (collectively referred to as Category D).

The evaluation criterion for these comparisons is "Quality." This was assessed using the LLM-based framework validated in Chapter 6, wherein the LLM evaluators (Claude-3-Opus and GPT-4o) were instructed to holistically assess poems on a 1-5 scale ("lowest quality" to "highest quality") without

predefined sub-criteria (details in Section 6.1). The resulting quality scores and rankings reflect each evaluator’s internal model of poetic quality, enabling direct comparison across human and AI categories.

To facilitate these comparisons, we constructed twelve distinct experimental conditions. Each condition involved one of the six AI-generated poem sets (Claude-3-Opus prompt-only, GPT-4o prompt-only, FT-GPT-4o-mini raw, FT-GPT-4o-mini refined, FT-GPT-4o raw, FT-GPT-4o refined) being evaluated alongside the three human poem categories (A, B, C). For every condition, a comparative dataset was formed, comprising 30 poems from each of the four categories (A, B, C, and the specific D set). These 120 poems were then evaluated by both Claude-3-Opus and GPT-4o independently, using 150 randomly drawn subsets of 16 poems (4 from each category per subset). This process yielded 12 sets of average quality scores for each of the four categories under comparison.

The overall mean quality rankings are visualized in Figure 19, and the detailed mean scores (Mean \pm SD) alongside ANOVA results for all 12 experimental conditions are presented in Table 25.

Evaluations by Claude-3-Opus						
LLM Poem Source (D)	Good (A)	Medium (B)	Bad (C)	LLM (D)	F-value	p-value
Claude-3-Opus	10.08 \pm 2.63	7.84 \pm 2.92	3.40 \pm 1.54	12.68 \pm 1.38	90.81	2.629×10^{-30}
GPT-4o	10.23 \pm 2.55	7.89 \pm 2.76	3.28 \pm 1.45	12.60 \pm 1.33	103.39	1.234×10^{-32}
FT-GPT-4o-mini (Raw)	12.79 \pm 2.27	9.76 \pm 2.71	3.70 \pm 1.79	7.75 \pm 2.26	82.57	1.173×10^{-28}
FT-GPT-4o-mini (Refined)	12.29 \pm 2.49	9.08 \pm 3.05	3.46 \pm 1.61	9.17 \pm 1.83	73.91	8.485×10^{-27}
FT-GPT-4o (Raw)	12.85 \pm 2.22	10.22 \pm 2.91	3.91 \pm 1.73	7.03 \pm 1.84	86.81	1.616×10^{-29}
FT-GPT-4o (Refined)	12.45 \pm 2.44	9.53 \pm 3.12	3.44 \pm 1.63	8.57 \pm 1.17	84.37	5.018×10^{-29}
Evaluations by GPT-4o						
LLM Poem Source (D)	Good (A)	Medium (B)	Bad (C)	LLM (D)	F-value	p-value
Claude-3-Opus	12.20 \pm 2.36	9.64 \pm 2.78	3.37 \pm 1.37	8.71 \pm 1.38	95.69	3.093×10^{-31}
GPT-4o	11.11 \pm 2.55	8.42 \pm 2.40	3.32 \pm 1.17	11.06 \pm 1.10	104.93	6.622×10^{-33}
FT-GPT-4o-mini (Raw)	12.36 \pm 2.16	9.73 \pm 2.41	3.87 \pm 1.41	7.96 \pm 2.17	87.72	1.063×10^{-29}
FT-GPT-4o-mini (Refined)	11.76 \pm 2.52	9.17 \pm 2.77	3.53 \pm 1.28	9.49 \pm 1.80	75.70	3.409×10^{-27}
FT-GPT-4o (Raw)	12.35 \pm 2.10	10.02 \pm 2.63	3.73 \pm 1.29	7.84 \pm 1.70	97.20	1.625×10^{-31}
FT-GPT-4o (Refined)	11.67 \pm 2.53	9.63 \pm 2.56	3.25 \pm 1.21	9.45 \pm 1.15	99.38	6.477×10^{-32}

Table 25: ANOVA Results: Comparison of Mean Quality Scores between Human Poem Categories (A, B, C) and LLM-Generated Poems (D) for Section 7.4 Experiments. Scores are Mean \pm STD.

General Observations and Statistical Significance

Initial inspection of Figure 19 and Table 25 reveals consistent patterns. Both LLM evaluators reliably distinguished between the human poetry tiers, generally ranking Human “Good” (A) above “Medium” (B), and “Medium” (B) above “Bad” (C). AI-generated poems (D), particularly those from prompt-only methods, achieved high quality ratings, often comparable to or exceeding Human “Good” (A) poems according to Claude-3-Opus, while GPT-4o rated Human “Good” (A) slightly higher or comparable to prompt-only AI. Fine-tuned AI poems (D), while still outperforming Human “Bad” (C), typically ranked

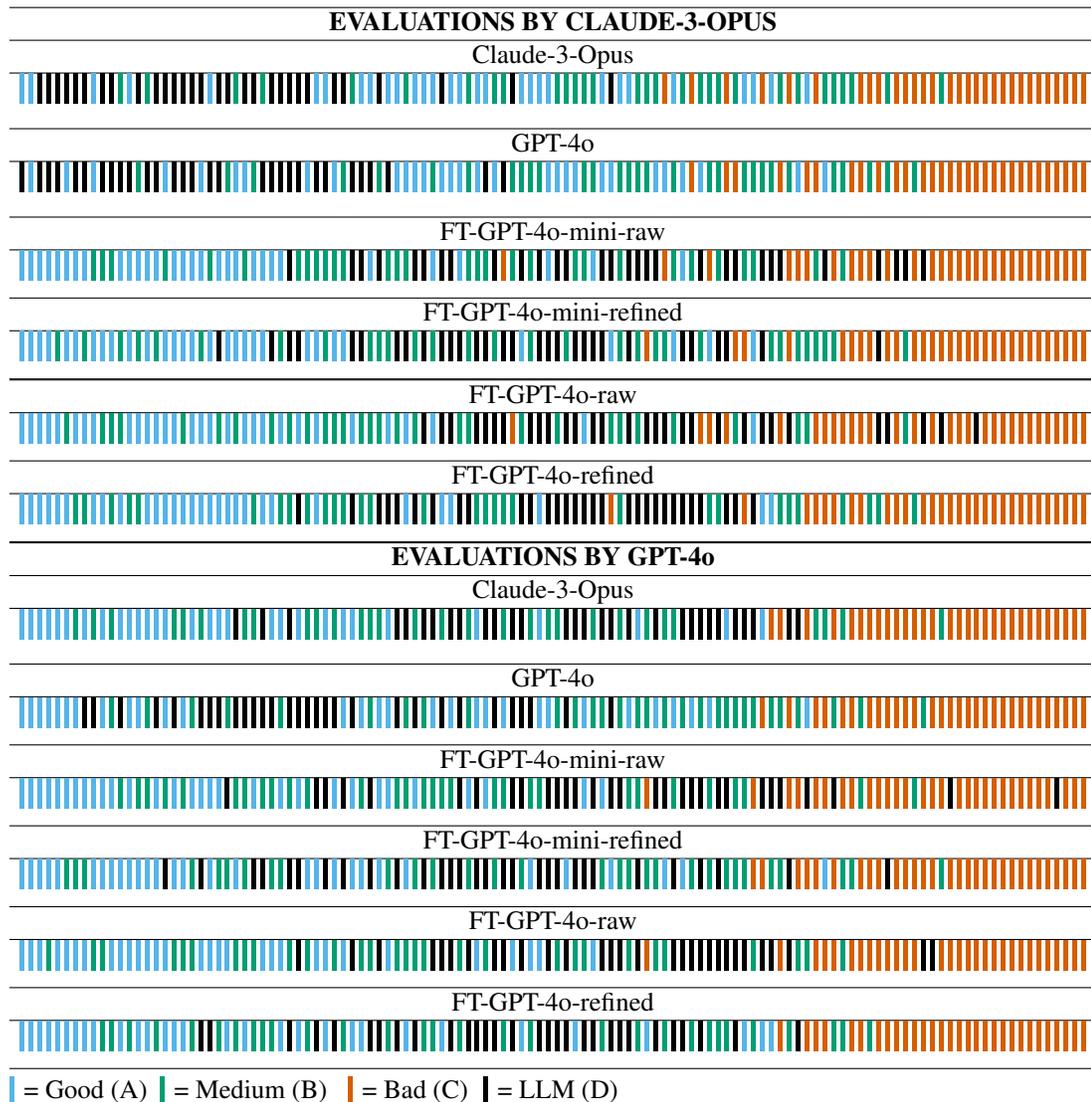


Figure 19: Ordering of the LLM-generated poems vs. human poems of three categories. LLM poetry generated by Claude and GPT-4o, evaluation done by Claude and GPT-4o.

below prompt-only AI (D) and Human “Good” (A) poems. A notable divergence was also apparent in the evaluators’ preferences, with Claude-3-Opus rating AI-generated poems (D) more favorably relative to human poems (A) than GPT-4o did, particularly for prompt-only AI poems.

To formally assess whether the observed differences in mean quality scores among the four poem categories (Human Good A, Human Medium B, Human Bad C, and the specific LLM-generated D set) were statistically significant, one-way Analysis of Variance (ANOVA) tests were conducted for each of the 12 experimental conditions. The null hypothesis (H_0) for each ANOVA was that the mean quality scores of all four categories were equal (e.g., $\mu_A = \mu_B = \mu_C = \mu_D$). The alternative hypothesis (H_1) was

that at least one category mean was different from the others. As Table 25 shows, the overall ANOVA was highly significant in every case (e.g., $F(3 \text{ groups} - 1, 120 \text{ total observations} - 4 \text{ groups}) = F(3, 116) \geq 73.91$, all $p\text{-values} \leq 8.485 \times 10^{-27}$, leading to the rejection of the null hypothesis in all conditions and indicating substantial variations in perceived quality across the poem categories. To identify which specific categories differed significantly from each other, Tukey's Honestly Significant Difference (HSD) post-hoc tests were performed ($\alpha = 0.05$). The key pairwise comparison results from these Tukey HSD tests are detailed below (adjusted p -values reported; $p = 0.000$ indicates $p < .001$ unless a more precise small p -value is given by Tukey's test).

Evaluations by Claude-3-Opus (Table 25, top panel):

- Poems from prompt-only **Claude-3-Opus (D)** ($M=12.68$) were rated significantly higher than Human Good (A) ($M=10.08$; Tukey HSD $p = 0.0004$). Similarly, poems from prompt-only **GPT-4o (D)** ($M=12.60$) were rated significantly higher than Human Good (A) ($M=10.23$; Tukey HSD $p < .001$). In both these conditions, Human Good (A) significantly outranked Human Medium (B) ($M=7.84$ and $M=7.89$ respectively; $p \leq 0.0011$), which in turn significantly outranked Human Bad (C) ($M=3.40$ and $M=3.28$ respectively; $p < .001$).
- For fine-tuned models, Human Good (A) poems ($M \approx 12.3$ - 12.9) were consistently rated significantly higher than their corresponding AI (D) counterparts: **FT-GPT-4o-mini (Raw)** ($M=7.75$, Tukey HSD $p < .001$), **FT-GPT-4o-mini (Refined)** ($M=9.17$, Tukey HSD $p < .001$), **FT-GPT-4o (Raw)** ($M=7.03$, Tukey HSD $p < .001$), and **FT-GPT-4o (Refined)** ($M=8.57$, Tukey HSD $p < .001$). In these fine-tuned conditions, Human Good (A) also significantly outranked Human Medium (B) (Tukey HSD $p < .001$), and Human Medium (B) significantly outranked Human Bad (C) (Tukey HSD $p < .001$).

Evaluations by GPT-4o (Table 25, bottom panel):

- When GPT-4o evaluated prompt-only poems from **Claude-3-Opus (D)**, Human Good (A) poems ($M=12.20$) were rated significantly higher than these Claude-3-Opus (D) poems ($M=8.71$, Tukey HSD $p < .001$). In this scenario, Human Good (A) also significantly outranked Human Medium (B) ($M=9.64$, Tukey HSD $p = 0.0001$). Human Medium (B) poems ($M=9.64$) were rated higher than Claude-3-Opus (D) poems ($M=8.71$), though this difference was not statistically significant (Tukey HSD $p = 0.2306$).
- However, when GPT-4o evaluated its own prompt-only poems (**GPT-4o (D)**), Human Good (A) poems ($M=11.11$) and these GPT-4o (D) poems ($M=11.06$) were not significantly different (Tukey HSD $p = 0.9947$). In this second scenario, both Human Good (A) and GPT-4o (D) poems significantly outranked Human Medium (B) ($M=8.42$, Tukey HSD $p < .001$ for both comparisons).

- In both these “prompt-only D” scenarios evaluated by GPT-4o, Human Medium (B) poems significantly outranked Human Bad (C) (M=3.37 and M=3.32 respectively; Tukey HSD $p < .001$ in both cases).
- Similar to Claude-3-Opus’s evaluations, Human Good (A) poems (M≈11.7-12.4) consistently outranked poems from the fine-tuned models: **FT-GPT-4o-mini (Raw)** (M=7.96, Tukey HSD $p < .001$), **FT-GPT-4o-mini (Refined)** (M=9.49, Tukey HSD $p < .001$), **FT-GPT-4o (Raw)** (M=7.84, Tukey HSD $p < .001$), and **FT-GPT-4o (Refined)** (M=9.45, Tukey HSD $p < .001$). Again, in these fine-tuned conditions, Human Good (A) also significantly outranked Human Medium (B) (Tukey HSD $p \leq 0.0002$), and Human Medium (B) significantly outranked Human Bad (C) (Tukey HSD $p < .001$).

These initial comparisons highlight the relative standings of different AI generation methods against human benchmarks. We now turn to a more focused analysis within the fine-tuned category itself: the specific impact of the automated refinement step.

Impact of the Automated Refinement Step on Fine-Tuned Poems

The data in Table 25 also allows for an examination of the automated refinement step’s impact (Section 7.2.3) on the perceived quality of poems from the fine-tuned models. We compared the mean scores assigned by each LLM evaluator to the “raw” versus the “refined” versions of these poems.

When **Claude-3-Opus evaluated FT-GPT-4o-mini** outputs, the refined version (M=9.17) scored higher than its raw counterpart (M=7.75). This refinement elevated its quality. While the raw version (D) was rated significantly below Human Medium (B) (M=9.76 for B vs M=7.75 for raw D, Tukey HSD $p = 0.0029$), the refined version’s score (D, M=9.17) was not significantly different from that of Human Medium (B) (M=9.08 for B vs M=9.17 for refined D, Tukey HSD $p = 1.0$, indicating no significant difference).

A similar improvement due to refinement was observed for **FT-GPT-4o evaluated by Claude-3-Opus**: the refined version (M=8.57) scored higher than the raw version (M=7.03). Again, this moved the fine-tuned poems (D) from being rated significantly below Human Medium (B) when raw (M=10.22 for B vs M=7.03 for raw D, Tukey HSD $p < .001$) to a level not significantly different from Human Medium (B) poems when refined (M=9.53 for B vs M=8.57 for refined D, Tukey HSD $p = 0.3713$).

When **GPT-4o evaluated FT-GPT-4o-mini**, refinement also led to an increase in quality, with the refined version (M=9.49) outscoring the raw version (M=7.96). This shifted its standing relative to Human Medium (B) poems; the raw version (D) was significantly lower (M=9.73 for B vs M=7.96 for raw D, Tukey HSD $p = 0.0024$), while the refined version (D, M=9.49) was statistically indistinguishable from Human Medium (B) (M=9.17 for B vs M=9.49 for refined D, Tukey HSD $p = 0.9595$).

The same pattern held for **FT-GPT-4o evaluated by GPT-4o**. The refined version (M=9.45) scored

higher than the raw version ($M=7.84$). Consequently, this refinement allowed the poems (D) to achieve comparability with Human Medium (B) poems ($M=9.63$ for B vs $M=9.45$ for refined D, Tukey HSD $p = 0.9942$), an improvement from the raw versions (D) which were rated significantly lower than Human Medium (B) ($M=10.02$ for B vs $M=7.84$ for raw D, Tukey HSD $p = 0.0003$).

Collectively, these comparisons of mean scores and significance testing reveal that the automated refinement step consistently improved the perceived quality of poems from both fine-tuned models, as judged by both LLM evaluators. While refinement did not elevate the fine-tuned poems to the level of Human “Good” (A) or the AI prompt-only outputs, it generally brought their quality from being rated significantly below Human “Medium” (B) to a level statistically comparable with Human “Medium” (B) poems. This suggests that refinement was effective in addressing superficial imperfections, thereby enhancing their overall quality assessment, an as such it deserves attention in future research.

Synthesis of Comparative Findings

These statistical analyses provide robust support for the initial observations. A clear hierarchy emerges: Human “Bad” (C) poems are consistently rated lowest. Human “Good” (A) poems are reliably identified as high quality. Prompt-only AI poems (D) achieve quality that is either comparable to (GPT-4o evaluator) or exceeds (Claude-3-Opus evaluator) Human “Good” (A) poems. Fine-tuned AI poems (D), even after refinement, generally align with or slightly surpass Human “Medium” (B) poems but do not reach the perceived quality of Human “Good” (A) or prompt-only AI outputs. Crucially, the divergence between Claude-3-Opus and GPT-4o in rating top-tier AI versus human poetry underscores a significant evaluator effect, with Claude-3-Opus showing a stronger preference for AI-generated content, especially from prompt-only methods.

Examples of specific highly-rated LLM poems from these evaluations are presented in Section 7.4.1.

7.4.1 Example poems from evaluations of AI-generated poems against human poetry

This section presents examples of highly-ranked LLM-generated poems alongside the top-ranked human poems from the evaluation scenarios detailed in Section 7.4. Each example includes the context of the evaluation (which LLM evaluated, which AI model generated the poem), the top human poems for comparison, and the highest-scoring LLM poem from that specific evaluation, along with its generating summary and with the details of the original poem from which the summary was generated. Scores are out of 16.

Claude-3-Opus evaluating Claude-3-Opus poems against human poems

In this evaluation, the first two top-scoring poems were from the Human “Good” category. The highest-scoring poem by Claude-3-Opus was in the third position.

1. Samiya Bashir “Carnot Cycle”, Category: Human “Good”, score: 15.76 (Bashir 2014)
2. Ocean Vuong “DetoNation”, Category: Human “Good”, score: 15.0 (Vuong 2013)
3. Claude-3-Opus “The Eternal Dance”, Category: LLM, score: 14.65

The Eternal Dance (by Claude-3-Opus)

Summary:

This poem explores the cyclical nature of existence, from cosmic creation to human life and death. It portrays a divine creator bringing worlds into being, and draws parallels between universal and human creation. The imagery of celestial bodies and earthly constructions emphasizes the grand scale of creation and the transience of human endeavors. The poem’s tone is contemplative and ultimately hopeful, suggesting that death is a return to the divine source and a reunion with loved ones. It presents life as a journey of continuous creation, destruction, and renewal, guided by a benevolent cosmic force.

Summary of: One By One by Ella Wheeler Wilcox.

Poem:

From stardust swirling in the cosmic dance,
 Worlds born anew at divine behest.
 Celestial bodies in eternal prance,
 A universe alive, forever blessed.

On Earth below, we mortals toil and strive,
 Our works mere echoes of creation grand.
 With fleeting joys and sorrows we contrive,
 To build our lives like castles in the sand.

Yet in the end, we all must yield to fate,
 Our bodies turned to dust from whence we came.
 The cycle spins with new lives to create,
 As we rejoin the source from which we came.

Though grief may cloud the heart with heavy shroud,
 Hope whispers soft of loved ones to embrace.
 In death’s release, the soul slips free and proud,

Returning home to its eternal place.
 So let us cherish life, each precious breath,
 While in this dance of birth and mortal death.

Claude-3-Opus evaluating GPT-4o poems against human poems

In this evaluation, the poem “The Turn of Seasons” written by GPT-4o was ranked the highest. The highest-scoring human poem was “Carnot Cycle” by Samiya Bashir, placed in the second position.

1. GPT-4o “The Turn of Seasons”, Category: LLM, score: 15.4
2. Samiya Bashir “Carnot Cycle”, Category: Human “Good”, score: 15.32 (Bashir 2014)

The Turn of Seasons (by GPT-4o)

Summary:

This allegorical poem depicts the transition from winter to spring through personification. King Winter, initially powerful and arrogant, is seduced by the Siren South Wind. Her flattery and affection gradually weaken him, allowing Spring to reclaim the land. The poem uses vivid imagery of icy landscapes and seasonal changes to symbolize the cyclical nature of time and the inevitable shift of power. The tone evolves from stern and domineering to tender and deceptive, ultimately ending with a sense of renewal. The central theme explores the impermanence of authority and the subtle yet potent force of persuasion in overcoming seemingly insurmountable obstacles.

Summary of: The King and the Siren by Ella Wheeler Wilcox.

Poem:

In realms where icy crowns command the night,
 King Winter roams with bitter, frost-bound breath,
 His cloak of snow reflects the pale moonlight,
 Dominion held through frozen, silent death.
 Yet whispers call from southern emerald seas,
 The Siren South Wind sings with silken tone,
 Her gentle voice, a mellowing, lulling breeze,
 Draws near the heart beneath the frost-bound throne.
 Through flattery and warmth, she breaks his will,
 Her soft caress thaws ice with tender wiles,
 The ground begins to tremble, quiet, still,

As Spring steps forth, with hopes and blooming smiles.

Now veils of white, once stubborn, fade away,

Revealing earth anew beneath the sun,

The reign of Winter yields to newborn day,

As life and light through every glade do run.

The tale unfolds, as time in cycles blend,

Where power wanes, another must ascend.

Claude-3-Opus evaluating FT-GPT-4o-mini poems against human poems

In this evaluation, the two highest-scoring poems were from the Human “Good” category. The highest-scoring poem produced by FT-GPT-4o-mini was ranked at position 23.

1. Samiya Bashir “Carnot Cycle”, Category: Human “Good”, score: 16.0 (Bashir 2014)
2. Aimee Nezhukumatathil “Two Moths”, Category: Human “Good”, score: 15.28 (Nezhukumatathil 2013)
- ...
23. FT-GPT-4o-mini “From Symphony to Silence: Love’s Cadence”, Category: LLM, score: 12.40

From Symphony to Silence: Love’s Cadence (by FT-GPT-4o-mini)

Summary:

This poem explores the transformative power of music and love on the human experience. The speaker describes being moved by various musical instruments, each evoking different emotions and insights into life’s complexities. The violin’s sorrow, the drum’s bravery, and the “mighty chords” all contribute to a rich tapestry of sensations. However, the poem culminates in the realization that the true symphony of life is found in the intimate connection of love, expressed through physical and emotional responses. The tone shifts from contemplative to passionate, emphasizing the profound impact of human connection over artistic expression.

Summary of: Music by Wilfred Owen.

Poem:

I was moved to hear the great violin
 Cry out, coiling like a serpent’s wake,
 And all unhappy malice that would break
 Love in a wrathful heart. When the din

Was at its height, it seemed to be within
 Some random destiny tumbling on our track.
 Questions would twist and twine and wrack,
 Coiling 'twixt the workers, toil, and scanty skin.

 I was moved to hear the great drum beat,
 Filling men's souls with hope and vanity,
 And all that satisfies our days' sole need:
 To kill, be adorned, and gluttoned be.
 Out of that dark chaos, one thing seemed most sweet:
 To die, as drums beat on, with chance to bleed.

 I have heard the mighty chords unfold:
 Their intertwisting glories swept my mind;
 But from the slow threshold of the door,
 Your anger and my wit, our laughter and love
 Choke out all music, and we now behold
 A language of lips, eyes, and fingers to explore.

Claude-3-Opus evaluating FT-GPT-4o poems against human poems

In this evaluation, the two highest-scoring poems were from the Human “Good” category. The highest-scoring poem produced by FT-GPT-4o was ranked at position 32.

1. Samiya Bashir “Carnot Cycle”, Category: Human “Good”, score: 15.79 (Bashir 2014)
2. Ocean Vuong “DetoNation”, Category: Human “Good”, score: 15.75 (Vuong 2013)
- ...
32. FT-GPT-4o “Love’s Thorn: A Poet’s Metamorphosis”, Category: LLM, score: 11.50

Love’s Thorn: A Poet’s Metamorphosis (by FT-GPT-4o)

Summary:

This poem explores the bittersweet transition from passionate love to estrangement. It portrays lovers moving from an idyllic, Eden-like state of intimacy to a muted, distant existence. The imagery shifts from warm, ethereal descriptions of spiritual connection to stark, cold images of bare trees. Key symbols include Eden, representing perfect love, and trees, symbolizing the lovers growing apart. The poem’s tone is melancholic and wistful, lamenting the loss of deep connection and the inability to recapture past

intimacy. It reflects on the ephemeral nature of intense love and the profound silence that follows its departure.

Summary of: *Wedded* by Isaac Rosenberg.

Poem:

Sweet Love, Sweet Thorn, when lightly to my heart
 I took your thrust, whereby I since am slain,
 And you through me forever must remain
 In English verse, first master of that art.

 Wherein I've imagined, in times now past,
 With mightier others who sit hushed, you deign
 A yet unfinished, pettish muse to sustain
 Like an old moth, pinioned and holding fast

 To silk wings in its yesteryear's attire;
 And though through new abandonment, or born
 Of summer's shifting lights, this failed dreamer
 May never now your tremulous cloak retie,
 Still—flower, I love you, cake of sacred fire,
 Until I truly learn, Love, Sweet Thorn.

GPT-4o evaluating Claude-3-Opus poems against human poems

In this evaluation, the two highest-scoring poems were from the Human “Good” category. The highest-scoring poem produced by Claude-3-Opus was ranked at position 25.

1. Ocean Vuong “DetoNation”, Category: Human “Good”, score: 15.8 (Vuong 2013)
2. Phillip B. Williams “Of Darker Ceremonies”, Category: Human “Good”, score: 15.45 (Williams 2013)
- ...
25. Claude-3-Opus “Timeless Dance at Cashel Rock”, Category: LLM, score: 11.95

Timeless Dance at Cashel Rock (by Claude-3-Opus)

Summary:

This poem explores the interplay between reality and imagination, set against the backdrop of Cashel Rock in Ireland. It delves into themes of consciousness, spirituality, and the nature of existence. The

poet juxtaposes ancient symbols (Sphinx, Buddha) with a dancing girl, representing the tension between intellect, spirituality, and physical perfection. The imagery shifts between concrete and abstract, blurring the lines between the mind's eye and tangible reality. The tone is contemplative and somewhat melancholic, reflecting on the constraints of human existence and the transformative power of art and imagination. Ultimately, the poem suggests that moments of profound insight can transcend time and physical limitations.

Summary of: *The Double Vision Of Michael Robartes* by William Butler Yeats.

Poem:

Upon Cashel Rock, a timeless stage unfolds
Where Sphinx and Buddha meet, and stories old
Entwine with visions of a dancing maid,
Her graceful form in stone and shadow played.

The mind's eye blurs the line twixt earth and dream,
As ancient wisdom flows in silent stream
Through weathered stone and moss-clad castle walls,
Where once kings reigned in long-forgotten halls.

In fleeting glimpse, the girl pirouettes by,
A symbol of perfection, never shy.
She leaps between the real and the conceived,
Imagination's child, to art bequeathed.

The weight of ages presses like a vise,
Yet in this moment, time is carved in ice.
The boundaries of flesh begin to fade
As inspiration's spark ignites the glade.

What seems eternal shifts like windblown sand,
Recast by artist's eye and poet's hand.
Though anchored firm in Ireland's verdant soil,
The soul takes flight, released from mortal coil.

GPT-4o evaluating GPT-4o poems against human poems

In this evaluation, the two highest-scoring poems were from the Human "Good" category. The highest-scoring poem produced by GPT-4o was ranked at position 8.

1. Ocean Vuong “DetoNation”, Category: Human “Good”, score: 15.56 (Vuong 2013)
2. Phillip B. Williams “Of Darker Ceremonies”, Category: Human “Good”, score: 15.2 (Williams 2013)
- ...
8. GPT-4o “The Veiled Forest”, Category: LLM, score: 13.3

The Veiled Forest (by GPT-4o)

Summary:

This poem explores the theme of hidden depths and unseen worlds within nature. It contrasts the visible beauty of the forest—stars, foxgloves, and marguerites—with the mysterious, unseen multitudes dwelling within. The imagery of light and darkness, along with the inaccessible interior of the forest, creates a sense of wonder and enigma. The poem suggests that despite the outward beauty we can perceive, there are always deeper, unreachable realms beyond our understanding. The tone is one of awe and contemplation, inviting readers to consider the limitations of human perception and the vast, unknowable aspects of the natural world.

Summary of: The Dark Forest by Edward Thomas.

Poem:

In the whispering woods where foxgloves nod,
 And stars like scattered diamonds grace the night,
 A deeper world resides beneath the sod,
 Cloaked in shadows, veiled from human sight.

Though marguerites adorn the forest floor,
 And moonlight bathes each leaf in silver glow,
 There lies a realm unseen, an endless lore—
 An echo of the mysteries we do not know.

The boughs conceal their secrets well,
 With roots entwined in earth’s embrace,
 Where silent multitudes of beings dwell,
 In corridors of time and sacred space.

Thus, stand we awed, on nature’s edge,
 Peering through the glimmering veil;
 Aware of every timeless pledge,
 Yet humbled by the untold tale.

GPT-4o evaluating FT-GPT-4o-mini poems against human poems

In this evaluation, the two highest-scoring poems were from the Human “Good” category. The highest-scoring poem produced by FT-GPT-4o-mini was ranked at position 17.

1. Ocean Vuong “DetoNation”, Category: Human “Good”, score: 15.76 (Vuong 2013)
2. Phillip B. Williams “Of Darker Ceremonies”, Category: Human “Good”, score: 15.36 (Williams 2013)
- ...
17. FT-GPT-4o-mini “Tithonus: Sentinel of Endless Dawn”, Category: LLM, score: 12.74

Tithonus: Sentinel of Endless Dawn (by FT-GPT-4o-mini)**Summary:**

This poem explores the bittersweet fate of Tithonus, granted immortality but not eternal youth. It portrays his aging and retreat to Aurora’s palace, where he lingers in a dreamlike state. The poem juxtaposes the decay of mortal life with the unchanging nature of immortality, using imagery of seasons, celestial bodies, and the sea. The tone is melancholic and contemplative, emphasizing the isolation and stagnation of endless existence. Through Tithonus’ eternal vigil, the poem reflects on the passage of time, memory, and the human condition, suggesting that immortality without youth is a burden rather than a blessing.

Summary of: Tithonus by Alan Seeger.

Poem:

Tithonus, granted immortality by Zeus,
 Ages endlessly, his mortal soul still bound.
 From Olympus’ heights, he now views
 Aurora’s daily rise, his fate profound.
 A sentinel of dawn, he sits enthroned,
 Keen-eyed, yet weary, sculpted by time,
 In an endless sea of watch, bemoaned,
 A drudge behind the Gods sublime.
 The poet, priest of laurel, recalls
 This figure in his verse, old-wakes
 Echoing through celestial halls,
 An unhopeful worship he makes.

GPT-4o evaluating FT-GPT-4o poems against human poems

In this evaluation, the two highest-scoring poems were from the Human “Good” category. The highest-scoring poem produced by FT-GPT-4o was ranked at position 21.

1. Ocean Vuong “DetoNation”, Category: Human “Good”, score: 15.95 (Vuong 2013)
2. Phillip B. Williams “Of Darker Ceremonies”, Category: Human “Good”, score: 15.41 (Williams 2013)
- ...
21. FT-GPT-4o “Foxglove Requiem in a Blood-Red West”, Category: LLM, score: 12.05

Foxglove Requiem in a Blood-Red West (by FT-GPT-4o)**Summary:**

This brief, evocative poem explores themes of love, desire, and spiritual connection through natural imagery. The speaker compares foxgloves at sunset to altar candles, creating a sacred atmosphere. The greenhouse roof serves as a metaphorical altar, suggesting elevation and transcendence. The poem’s central conceit links the act of lifting a loved one with understanding the flowers’ “burning”, implying that physical closeness leads to deeper emotional and spiritual insight. The tone is intimate and reverent, blending romantic longing with religious undertones. Through these vivid images, the poem captures a moment of intense feeling and potential transformation, where nature, love, and spirituality intertwine.

Summary of: Vespers by Amy Lowell.

Poem:

From where I stand, the foxgloves look like altar candles,
 And the greenhouse roof is the church roof,
 And the damp weeds are the people.
 I stand on the road and lift the one I love
 Against the blood-red west.
 Already all the women are in black.
 Tomorrow it will rain on the church roof
 And on the stones and on the weeds.
 And I shall know
 The flowers are burning in the fox-beds,
 And I shall not wonder
 That I am where the burning is.

I shall watch the candles gleaming
Round her forehead,
And no guiding touch from shoulder or from hand
Will shudder me back to smoldering understanding
Too dark to bear.

7.5 Comparing LLM-generated Poems Across Models

Having compared AI-generated poetry to human benchmarks, we now turn to comparing the *quality* of poems generated using different AI methods against each other. Beyond establishing a relative quality hierarchy among these AI approaches, the experiments in this section also serve to further verify our LLM evaluators' ability to discern and consistently rank varying levels of poetic quality, reinforcing the findings from the previous comparisons.

7.5.1 Evaluating Four AI Generation Methods

To compare the overall quality achieved by each of the four AI generation methods (Claude-3-Opus prompt-only, GPT-4o prompt-only, fine-tuned GPT-4o-mini, fine-tuned GPT-4o), we assembled a combined dataset. This dataset consisted of the 30 highest-quality poems selected from each method, as determined by the process in Section 7.3, resulting in a total of 120 AI-generated poems. We then generated 150 evaluation subsets, each containing 16 poems (4 randomly selected poems from each of the four generation methods), shuffling the order within each subset.

These subsets were evaluated for “Quality” by both Claude-3-Opus and GPT-4o using our established in-context ranking methodology (Chapter 6), with rank position determining the score. Figure 20 presents a visual representation of the resulting mean quality rankings. The detailed statistical analysis, including mean scores (Mean \pm SD) and ANOVA results for each of the four experimental conditions (varying the evaluator and whether raw or refined versions of fine-tuned poems were used), is presented in Table 26.

One-way ANOVAs were conducted for each of the four experimental conditions to compare the mean quality scores across the four AI generation methods. As shown in Table 26, the overall ANOVA was highly significant for every condition:

- Claude-3-Opus evaluation with refined FT poems: $F(3, 116) = 122.80, p = 7.51 \times 10^{-36}$
- GPT-4o evaluation with refined FT poems: $F(3, 116) = 42.76, p = 1.10 \times 10^{-18}$
- Claude-3-Opus evaluation with raw FT poems: $F(3, 116) = 217.77, p = 1.77 \times 10^{-47}$
- GPT-4o evaluation with raw FT poems: $F(3, 116) = 100.68, p = 3.77 \times 10^{-32}$

Evaluation (Condition)	Claude-3-Opus (Prompt)	GPT-4o (Prompt)	GPT-4o-mini (Fine-tuned)	GPT-4o (Fine-tuned)	F-value	p-value
Mean Quality Scores \pm SD						
Claude (Raw)	12.09 \pm 1.55	12.38 \pm 1.58	4.68 \pm 1.41	4.84 \pm 1.74	217.77	1.77×10^{-47}
Claude (Ref.)	11.31 \pm 1.80	11.85 \pm 1.83	5.82 \pm 1.80	5.02 \pm 1.56	122.80	7.51×10^{-36}
GPT-4o (Raw)	10.77 \pm 1.61	11.93 \pm 1.49	5.40 \pm 1.88	5.88 \pm 2.05	100.68	3.77×10^{-32}
GPT-4o (Ref.)	9.34 \pm 1.94	10.99 \pm 1.57	6.89 \pm 1.69	6.78 \pm 1.57	42.76	1.10×10^{-18}

Table 26: ANOVA Results: Comparison of Mean Quality Scores among Four AI Generation Methods (Section 7.5.1). Each row represents a separate ANOVA comparing the four AI methods listed as columns, evaluated by the specified LLM and using either refined or raw versions of fine-tuned poems. Degrees of freedom for F-statistic are (3, 116) for all tests. The “Ref.” or “Raw” in the “Evaluation (Condition)” column indicates whether the refined or raw versions of the fine-tuned poems were included in that specific 4-way comparison.

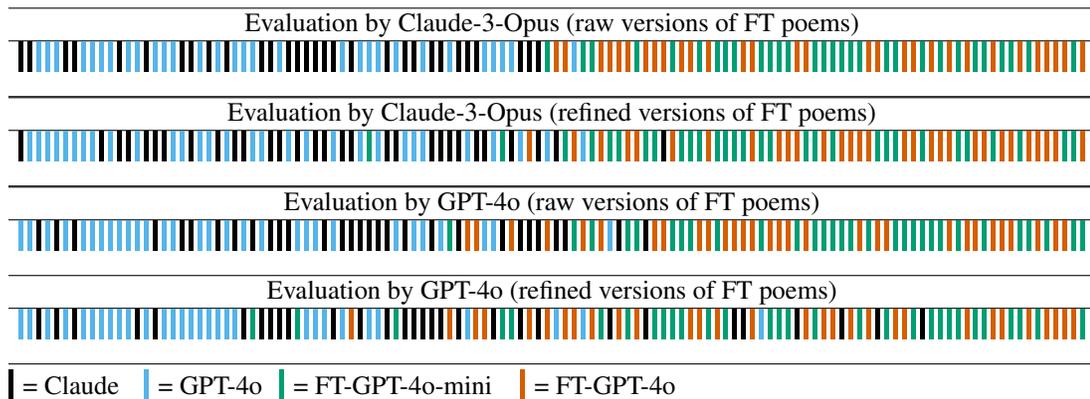


Figure 20: Ordering of the LLM-generated poems by Claude, GPT-4o, FT-GPT-4o and FT-GPT-4o-mini.

These results indicate substantial differences in perceived quality among the AI methods. Subsequent Tukey’s HSD post-hoc comparisons ($\alpha = 0.05$) identified the specific pairwise differences, with key findings (adjusted p-values) detailed below.

When Raw (Unrefined) Fine-Tuned Poems were Used in Comparison:

- **Evaluated by Claude-3-Opus** ($F(3, 116) = 217.77$, $p = 1.77 \times 10^{-47}$): Poems from Claude-3-Opus (Prompt) ($M=12.09$) and GPT-4o (Prompt) ($M=12.38$) were not significantly different from each other (Tukey HSD $p = 0.8594$). However, both prompt-based methods were rated significantly higher than FT-GPT-4o (Raw) ($M=4.84$) and FT-GPT-4o-mini (Raw) ($M=4.68$) (all relevant Tukey HSD $p = 0.0$). The two raw fine-tuned methods were not significantly different from each other (Tukey HSD $p = 0.9821$).
- **Evaluated by GPT-4o** ($F(3, 116) = 100.68$, $p = 3.77 \times 10^{-32}$): Poems from GPT-4o (Prompt)

(M=11.93) and Claude-3-Opus (Prompt) (M=10.77) were not significantly different (Tukey HSD $p = 0.0826$). Both prompt-based methods significantly outscored FT-GPT-4o (Raw) (M=5.88) and FT-GPT-4o-mini (Raw) (M=5.40) (all relevant Tukey HSD $p = 0.0$). The two raw fine-tuned methods were not significantly different from each other (Tukey HSD $p = 0.7203$).

When Refined Fine-Tuned Poems were Used in Comparison:

- **Evaluated by Claude-3-Opus** ($F(3, 116) = 122.80, p = 7.51 \times 10^{-36}$): Poems from GPT-4o (Prompt) (M=11.85) and Claude-3-Opus (Prompt) (M=11.31) were not significantly different from each other (Tukey HSD $p = 0.4836$). Both prompt-based methods were rated significantly higher than FT-GPT-4o-mini (Refined) (M=5.82) and FT-GPT-4o (Refined) (M=5.02) (all relevant Tukey HSD $p = 0.0$). The two refined fine-tuned methods were not significantly different from each other (Tukey HSD $p = 0.2912$).
- **Evaluated by GPT-4o** ($F(3, 116) = 42.76, p = 1.10 \times 10^{-18}$): GPT-4o (Prompt) poems (M=10.99) were rated significantly higher than Claude-3-Opus (Prompt) poems (M=9.34; Tukey HSD $p = 0.0018$). Both prompt-based methods significantly outscored FT-GPT-4o (Refined) (M=6.78) and FT-GPT-4o-mini (Refined) (M=6.89) (all relevant Tukey HSD $p = 0.0$). The two refined fine-tuned methods were not significantly different from each other (Tukey HSD $p = 0.9999$).

These statistical analyses confirm the descriptive trends observed in Figure 20 and the mean scores in Table 26. Across all four experimental conditions, a clear quality preference hierarchy emerges. Echoing the strong performance of prompt-based generation noted in Section 7.4, both Claude-3-Opus and GPT-4o evaluators consistently rated poems generated via direct prompting (Claude-3-Opus (Prompt) and GPT-4o (Prompt)) as significantly higher quality than those produced by either of the fine-tuned models (FT-GPT-4o-mini and FT-GPT-4o), regardless of whether the raw or refined versions of the fine-tuned poems were considered (all relevant Tukey HSD $p = 0.0$, indicating $p < .001$).

Contextualizing these findings requires acknowledging the refinement step's impact. Although automated refinement, as shown with statistical significance in Section 7.4 (Table 25), elevated the perceived quality of fine-tuned poems to a level often comparable with Human "Medium" poetry, and led to observable mean score improvements in the current AI-versus-AI comparisons (Table 26), this was insufficient to match the prompt-only AI methods. Indeed, even after refinement, both evaluators consistently rated the fine-tuned poems significantly lower than those produced by direct prompting.

Within the prompt-only generation methods, when Claude-3-Opus was the evaluator, there was no significant difference in perceived quality between poems generated by itself and those by GPT-4o (Tukey HSD $p = 0.8594$ for raw FT comparison; $p = 0.4836$ for refined FT comparison). However, when GPT-4o was the evaluator, it rated its own prompt-generated poems significantly higher than those

from Claude-3-Opus when refined FT poems were in the comparison set (Tukey HSD $p = 0.0018$); this difference was not significant when raw FT poems were in the set (Tukey HSD $p = 0.0826$). This again points to a potential evaluator effect or self-preference, particularly in the refined condition.

Among the fine-tuned generation methods (FT-GPT-4o vs. FT-GPT-4o-mini), there were no statistically significant differences in perceived quality in any of the four conditions (all relevant Tukey HSD p -values were high, e.g., $p = 0.9821$ for raw/Claude, $p = 0.7203$ for raw/GPT-4o, $p = 0.2912$ for refined/Claude, and $p = 0.9999$ for refined/GPT-4o).

The considerable disparity in average scores strongly suggests that, for achieving high perceived quality as judged by these LLMs, the general capabilities of untuned state-of-the-art models currently surpass the outputs from our fine-tuning approach in this general poetry generation task. Specifically, poems from fine-tuned models (with mean scores ranging from $M=4.68$ to $M=5.88$ for raw versions, and $M=5.02$ to $M=6.89$ for refined versions) scored significantly lower than those from prompt-based methods (mean scores ranging from $M=9.34$ to $M=12.38$, as detailed in Table 26).

This finding is particularly noteworthy. It suggests that while fine-tuning might be necessary for specific tasks like precise stylistic imitation (as indicated in Chapter 5), it does not necessarily lead to higher overall perceived quality compared to leveraging the broad capabilities of the base models through prompting for general poetry generation. The distinction between achieving specific constraints (like style) and maximizing general quality perception warrants further investigation in the context of evolving LLM architectures and training paradigms.

7.5.2 Highest scoring poems from evaluating four datasets of AI-generated poetry

This section presents examples of the highest-ranked LLM-generated poems from the direct comparison of the four AI generation methods, as detailed in Section 7.5.1. Each example includes the context of the evaluation (which LLM evaluated) and the top-scoring LLM poems from that specific evaluation, along with their generating summaries. The evaluation involved comparing poems from Claude-3-Opus (Prompt), GPT-4o (Prompt), FT-GPT-4o-mini (Fine-tuned), and FT-GPT-4o (Fine-tuned). All scores are out of 16.

Evaluations by Claude-3-Opus (refined versions of fine-tuned poems)

In the evaluations by Claude-3-Opus where the refined versions of the fine-tuned poems were used for comparison, the highest scoring poem was “The Eternal Dance” written by Claude-3-Opus (Prompt), and the second highest was “The Veiled Forest” by GPT-4o (Prompt).

1. “The Eternal Dance”, Score: 14.89, Generated by: Claude-3-Opus (Prompt)

2. “The Veiled Forest”, Score: 14.69, Generated by: GPT-4o (Prompt)

Both of these poems were previously presented in Section 7.4.1.

Evaluations by GPT-4o (refined versions of fine-tuned poems)

In the evaluations by GPT-4o where the refined versions of the fine-tuned poems were used for comparison, the highest scoring poem was “The Veiled Forest” written by GPT-4o (Prompt), and the second highest was “Grace in the Ordinary” also by GPT-4o (Prompt).

1. “The Veiled Forest”, Score: 13.81, Generated by: GPT-4o (Prompt)
2. “Grace in the Ordinary”, Score: 13.64, Generated by: GPT-4o (Prompt)

“The Veiled Forest” by was presented in section 7.4.1.

Grace in the Ordinary (by GPT-4o)

Summary:

This poem celebrates the beauty and significance of everyday objects and sensations. The central theme is the appreciation of “common things”, from household items to natural elements. The poet uses vivid imagery to evoke the tactile and sensory experiences associated with these objects, such as the “touch, feel, smell of paper” and the “touch of clouds”. The poem’s tone is one of reverence and wonder, finding grace in the ordinary. By highlighting the “deariness” of these items, the poet encourages readers to find joy and meaning in the simple aspects of life often taken for granted.

Summary of: Common Things by Ivor Gurney.

Poem:

In the hum of a morning light, soft and clear,
Lies the quiet glory of things we hold dear.
The kettle’s song as it starts to steam,
The dance of dust in a sunbeam.

Wooden floor warmed by bare, waking feet,
Whispers of stories in each creak and creak.
The touch, feel, smell of a page well-turned,
Libraries of wonder where dreams are learned.

Petals surrendering to the morning dew,
Clouds like cotton, vast and new.
Scribbles on paper, a child’s own art,

A universe captured in every heart.
 The crisp, fresh air of a walk alone,
 Echoes of footsteps where silence is grown.
 To see the unseen in the common place,
 Is to find the profound in a simple grace.

7.5.3 Evaluating Top vs. Bottom Poems Across Methods

As a final validation of our evaluation methodology’s internal consistency for quality assessment, we conducted an experiment comparing the extremes of perceived quality within our AI-generated datasets. We selected the two highest-ranked (“top”) and two lowest-ranked (“bottom”) poems from each of the four AI generation methods (Claude-3-Opus prompt-only, GPT-4o prompt-only, fine-tuned GPT-4o, and fine-tuned GPT-4o-mini). This selection was based on the initial quality screening detailed in Section 7.3, where fine-tuned poems were assessed in their “raw” state. This process created a focused set of 16 poems (8 “top” and 8 “bottom”) representing the quality spectrum across all AI methods. This combined set of 16 poems was then re-evaluated under four distinct conditions: using either Claude-3-Opus or GPT-4o as the evaluator, and by considering the fine-tuned poems within the set as either their original “raw” versions or their “refined” versions (produced as per Section 7.2.3). For each of these four re-evaluation scenarios, 30 evaluation subsets were generated, each containing all 16 selected poems in a randomized order. These subsets were then evaluated for “Quality” by the designated LLM, with scores derived from the poem’s rank position within each subset.

To statistically assess whether poems initially identified as “top” quality were indeed rated significantly higher by the LLM evaluators than those identified as “bottom” quality in these new mixed assessments, independent samples Welch’s t-tests were conducted for each of the four evaluation conditions. The resulting average quality scores (Mean \pm SD) for the “top” (N=8) and “bottom” (N=8) groups of poems, along with the t-test results, are presented in Table 27.

The data presented in Table 27 clearly demonstrates the LLM evaluators’ effectiveness in discerning quality differences between these pre-selected groups. In all four conditions, poems initially identified as “top” quality received significantly higher average scores than those identified as “bottom” quality:

- When **Claude-3-Opus evaluated the set including raw FT versions**, “top” poems (M=12.13, SD=2.28) were rated significantly higher than “bottom” poems (M=4.87, SD=2.55) (Welch’s $t \approx 6.01$, $p = 3.390 \times 10^{-5}$).
- When **Claude-3-Opus evaluated the set including refined FT versions**, “top” poems (M=12.33, SD=1.21) again significantly outscored “bottom” poems (M=4.68, SD=2.01) (Welch’s $t \approx 9.23$,

Evaluation Condition	M Top (SD)	M Bottom (SD)	t-stat	p-value
Claude Eval (Raw FT versions)	12.13 (2.28)	4.87 (2.55)	6.01	3.390×10^{-5}
Claude Eval (Refined FT versions)	12.33 (1.21)	4.68 (2.01)	9.23	1.191×10^{-6}
GPT-4o Eval (Raw FT versions)	10.95 (3.57)	6.02 (1.17)	3.71	5.387×10^{-3}
GPT-4o Eval (Refined FT versions)	11.18 (2.89)	5.82 (2.26)	4.13	1.139×10^{-3}

Table 27: T-test Results: Comparison of Mean Quality Scores between pre-selected “Top” (N=8) and “Bottom” (N=8) AI-Generated Poems (Section 7.3). “M Top (SD)” refers to Mean (Standard Deviation) for the “Top” group; “M Bottom (SD)” for the “Bottom” group. The “(Raw FT versions)” or “(Refined FT versions)” indicates the status of fine-tuned poems in the set. Degrees of freedom for t-tests are ≈ 14 (Welch’s t-test).

$$p = 1.191 \times 10^{-6}.$$

- When **GPT-4o evaluated the set including raw FT versions**, “top” poems (M=10.95, SD=3.57) were rated significantly higher as a group than “bottom” poems (M=6.02, SD=1.17) (Welch’s $t \approx 3.71$, $p = 5.387 \times 10^{-3}$). This overall group distinction held despite instances where individual poems pre-selected as “top” by their original evaluator received lower scores from GPT-4o; for example, “Eternal Faces of Beloved: A Devotion” (from FT-GPT-4o-mini, raw) scored M=5.00 in this condition (details in Table 28).
- This significant difference persisted when **GPT-4o evaluated the set including refined FT versions**: “top” poems (M=11.18, SD=2.89) significantly outscored “bottom” poems (M=5.82, SD=2.26) (Welch’s $t \approx 4.13$, $p = 1.139 \times 10^{-3}$). Again, some poems initially ranked as “top” received divergent scores; notably, “The Scarlet Shield” (from Claude-3-Opus, TOP) scored M=7.77, and the refined version of “Eternal Faces of Beloved: A Devotion” scored M=6.77 under GPT-4o’s evaluation in this condition (Table 29).

This consistent internal differentiation between poems pre-identified as higher and lower quality during our initial screening (Section 7.3) reinforces confidence in the LLM evaluation approach used throughout this chapter. It validates that the quality distinctions observed in the preceding comparisons (Sections 7.4 and 7.5) are based on stable and discernible quality judgments made by the LLM evaluators. The robustness of this differentiation is evident even when occasional individual poem scores deviate from their initial group selection. Such deviations, like the lower scores assigned by GPT-4o to “Eternal Faces of Beloved: A Devotion” and “The Scarlet Shield” (both pre-selected as “top” by other systems), did not disrupt the overall significant difference between the ‘top’ and ‘bottom’ group means and further highlight the evaluator-specific assessment patterns discussed earlier. Furthermore, inspection of the individual poem scores within the “top” group (presented in Table 28 and Table 29) reaffirms that poems generated via prompting generally rank among the highest quality outputs, even within this “top” tier.

Evaluation by Claude-3-Opus		
Poem Title	Avg. Score \pm STD	Model
The Mountain Endures	14.63 \pm 1.82	Claude-3-Opus TOP
The Scarlet Shield	14.50 \pm 1.77	Claude-3-Opus TOP
Hope’s Enduring Flight	14.30 \pm 1.04	GPT-4o TOP
Grace in the Ordinary	12.83 \pm 1.42	GPT-4o TOP
From Golden Walls to Rooted Prayers	11.30 \pm 1.79	FT-GPT-4o-mini TOP
Luminous Sacrament: Love’s Eternal Dawn	10.67 \pm 1.74	FT-GPT-4o-big TOP
Eternal Echoes: Love’s Resurrection Light	10.37 \pm 2.26	FT-GPT-4o-big TOP
<i>The Whirl of Thoughts</i>	<i>9.23 \pm 2.25</i>	<i>GPT-4o BOTTOM</i>
Eternal Faces of Beloved: A Devotion	8.47 \pm 2.11	FT-GPT-4o-mini TOP
<i>Echoes of Freedom in Captive Hearts</i>	<i>6.83 \pm 2.38</i>	<i>FT-GPT-4o-big BOTTOM</i>
<i>Love’s Unbalanced Scales: Wide World Over</i>	<i>5.97 \pm 1.17</i>	<i>FT-GPT-4o-mini BOTTOM</i>
<i>Maiden to Wife: Love’s Uncharted Voyage</i>	<i>5.90 \pm 1.58</i>	<i>FT-GPT-4o-mini BOTTOM</i>
<i>Whispers from a Distant Star</i>	<i>3.93 \pm 1.12</i>	<i>FT-GPT-4o-big BOTTOM</i>
<i>Ribbons of Silence</i>	<i>2.97 \pm 1.68</i>	<i>GPT-4o BOTTOM</i>
<i>The Persistent Peddlers</i>	<i>2.23 \pm 0.99</i>	<i>Claude-3-Opus BOTTOM</i>
<i>Five Finger Family</i>	<i>1.87 \pm 1.33</i>	<i>Claude-3-Opus BOTTOM</i>
Evaluation by GPT-4o		
Poem Title	Avg. Score \pm STD	Model
The Mountain Endures	15.13 \pm 0.96	Claude-3-Opus TOP
Grace in the Ordinary	14.50 \pm 1.43	GPT-4o TOP
Hope’s Enduring Flight	14.47 \pm 1.20	GPT-4o TOP
From Golden Walls to Rooted Prayers	10.81 \pm 2.29	FT-GPT-4o-mini TOP
The Scarlet Shield	10.07 \pm 3.69	Claude-3-Opus TOP
Luminous Sacrament: Love’s Eternal Dawn	9.73 \pm 3.66	FT-GPT-4o-big TOP
Eternal Echoes: Love’s Resurrection Light	7.90 \pm 2.79	FT-GPT-4o-big TOP
<i>Echoes of Freedom in Captive Hearts</i>	<i>7.83 \pm 3.75</i>	<i>FT-GPT-4o-big BOTTOM</i>
<i>Ribbons of Silence</i>	<i>7.23 \pm 3.68</i>	<i>GPT-4o BOTTOM</i>
<i>Love’s Unbalanced Scales: Wide World Over</i>	<i>6.33 \pm 3.59</i>	<i>FT-GPT-4o-mini BOTTOM</i>
<i>Whispers from a Distant Star</i>	<i>6.03 \pm 3.61</i>	<i>FT-GPT-4o-big BOTTOM</i>
<i>The Persistent Peddlers</i>	<i>5.87 \pm 3.01</i>	<i>Claude-3-Opus BOTTOM</i>
<i>The Whirl of Thoughts</i>	<i>5.80 \pm 2.87</i>	<i>GPT-4o BOTTOM</i>
Eternal Faces of Beloved: A Devotion	5.00 \pm 2.89	FT-GPT-4o-mini TOP
<i>Five Finger Family</i>	<i>4.93 \pm 2.94</i>	<i>Claude-3-Opus BOTTOM</i>
<i>Maiden to Wife: Love’s Uncharted Voyage</i>	<i>4.14 \pm 3.48</i>	<i>FT-GPT-4o-mini BOTTOM</i>

Table 28: LLM Poem Scores with updated data (raw versions from FT models). *Top-ranked entries are shown in bold, bottom-ranked entries in italics.*

Evaluation by Claude-3-Opus		
Poem Title	Avg. Score \pm STD	Model
The Mountain Endures	13.97 \pm 2.41	Claude-3-Opus TOP
The Scarlet Shield	13.80 \pm 3.15	Claude-3-Opus TOP
Hope’s Enduring Flight	13.07 \pm 1.79	GPT-4o TOP
Luminous Sacrament: Love’s Eternal Dawn	12.20 \pm 2.06	FT-GPT-4o-big TOP
Eternal Echoes: Love’s Resurrection Light	12.03 \pm 2.47	FT-GPT-4o-big TOP
From Golden Walls to Rooted Prayers	11.77 \pm 2.32	FT-GPT-4o-mini TOP
Grace in the Ordinary	11.20 \pm 1.90	GPT-4o TOP
Eternal Faces of Beloved: A Devotion	10.57 \pm 1.87	FT-GPT-4o-mini TOP
<i>Maiden to Wife: Love’s Uncharted Voyage</i>	<i>7.20 \pm 1.85</i>	<i>FT-GPT-4o-mini BOTTOM</i>
<i>The Whirl of Thoughts</i>	<i>6.17 \pm 1.44</i>	<i>GPT-4o BOTTOM</i>
<i>Echoes of Freedom in Captive Hearts</i>	<i>6.03 \pm 2.60</i>	<i>FT-GPT-4o-big BOTTOM</i>
<i>Love’s Unbalanced Scales: Wide World Over</i>	<i>5.67 \pm 1.40</i>	<i>FT-GPT-4o-mini BOTTOM</i>
<i>Whispers from a Distant Star</i>	<i>5.23 \pm 1.99</i>	<i>FT-GPT-4o-big BOTTOM</i>
<i>Ribbons of Silence</i>	<i>2.90 \pm 2.12</i>	<i>GPT-4o BOTTOM</i>
<i>The Persistent Peddlers</i>	<i>2.33 \pm 1.14</i>	<i>Claude-3-Opus BOTTOM</i>
<i>Five Finger Family</i>	<i>1.87 \pm 1.31</i>	<i>Claude-3-Opus BOTTOM</i>
Evaluation by GPT-4o		
Poem Title	Avg. Score \pm STD	Model
Hope’s Enduring Flight	14.57 \pm 1.69	GPT-4o TOP
Grace in the Ordinary	14.20 \pm 2.39	GPT-4o TOP
The Mountain Endures	13.33 \pm 2.88	Claude-3-Opus TOP
From Golden Walls to Rooted Prayers	11.67 \pm 2.21	FT-GPT-4o-mini TOP
Luminous Sacrament: Love’s Eternal Dawn	11.33 \pm 2.88	FT-GPT-4o-big TOP
Eternal Echoes: Love’s Resurrection Light	9.80 \pm 3.21	FT-GPT-4o-big TOP
<i>Love’s Unbalanced Scales: Wide World Over</i>	<i>8.23 \pm 2.64</i>	<i>FT-GPT-4o-mini BOTTOM</i>
<i>Whispers from a Distant Star</i>	<i>8.10 \pm 3.45</i>	<i>FT-GPT-4o-big BOTTOM</i>
<i>Echoes of Freedom in Captive Hearts</i>	<i>7.83 \pm 3.46</i>	<i>FT-GPT-4o-big BOTTOM</i>
The Scarlet Shield	7.77 \pm 3.90	Claude-3-Opus TOP
<i>Maiden to Wife: Love’s Uncharted Voyage</i>	<i>6.87 \pm 3.00</i>	<i>FT-GPT-4o-mini BOTTOM</i>
Eternal Faces of Beloved: A Devotion	6.77 \pm 2.88	FT-GPT-4o-mini TOP
<i>Ribbons of Silence</i>	<i>5.50 \pm 3.25</i>	<i>GPT-4o BOTTOM</i>
<i>The Whirl of Thoughts</i>	<i>4.20 \pm 1.92</i>	<i>GPT-4o BOTTOM</i>
<i>The Persistent Peddlers</i>	<i>3.00 \pm 2.14</i>	<i>Claude-3-Opus BOTTOM</i>
<i>Five Finger Family</i>	<i>2.83 \pm 2.90</i>	<i>Claude-3-Opus BOTTOM</i>

Table 29: LLM Poem Scores for top and bottom poems (refined versions from FT models). Top-ranked entries are shown in bold, bottom-ranked entries in italics.

The detailed scores for each of the 16 poems under the different evaluation conditions are presented below. Table 28 shows the results when raw versions of fine-tuned poems were part of the evaluated set, and Table 29 shows the results when refined versions were used. Section 7.5.4 gives the complete text of all 16 poems used in this experiment, along with the summaries they were generated from.

7.5.4 Poems from the Top vs. Bottom Quality Comparison

This section presents all 16 poems that were evaluated in the experiment comparing the pre-selected “top” and “bottom” quality AI-generated poems (the methodology for this experiment is detailed in Section 7.5.3). The poems produced by the FT-GPT-4o models and presented here are their “refined” versions. First, we present all the TOP poems, followed by all the BOTTOM poems, along with their generating summaries, together with the details of the original poem from which the summary was generated.

The Mountain Endures (by Claude-3-Opus) - TOP This poem was previously presented in full in Section 7.3.2.

The Scarlet Shield (by Claude-3-Opus) - TOP

Summary:

This poem explores divine judgment and redemption through Christ’s sacrifice. The speaker pleads with God to show mercy to sinners, repeatedly invoking Christ’s death as a shield between God’s wrath and humanity. Vivid imagery contrasts the “scarlet” of sin with the purifying power of Christ’s blood. The tone is both supplicatory and assertive, as the speaker reminds God of Christ’s ransom for humanity. The poem emphasizes the magnitude of human suffering and sin, while simultaneously highlighting the even greater power of Christ’s sacrifice. It portrays a complex relationship between God’s justice and mercy, mediated through Christ’s redemptive act.

Summary of: Mediation by Katharine Tynan.

Poem:

O Lord, behold the crimson stains that mar
 Our souls, the scarlet shame that we have wrought.
 The weight of sin, a crushing, leaden scar,
 Has left us lost, entangled and distraught.
 Yet in our darkest hour, a light divine
 Breaks through - the sacrifice of Christ, Thy Son.
 His blood, a holy, cleansing, healing wine;
 Through Him, the victory o’er death is won.

We plead, O God, let not Thy wrath ignite
 Against Thy children, though we merit blame.
 Remember Jesus' cross, that day of night;
 His anguish bore the depth of all our shame.

Have mercy, Lord! The Lamb has paid our debt.
 His love, a shield, holds back the fires of hell.
 The ransom's paid, Thy justice wholly met;
 In Christ alone, our hope and peace to dwell.

O wash us clean and make our garments white
 Our sins, though scarlet, banished by Thy Son.
 Redeemed, restored, brought back into the light,
 By grace through faith, the work of love is done.

Hope's Enduring Flight (by GPT-4o) - TOP This poem was previously presented in full in Section 7.3.2.

Grace in the Ordinary (by GPT-4o) - TOP This poem was previously presented in full in Section 7.5.2.

From Golden Walls to Rooted Prayers (by FT-GPT-4o-mini) - TOP This poem was previously presented in full in Section 7.3.2.

Eternal Faces of Beloved: A Devotion (by FT-GPT-4o-mini) - TOP

Summary:

This poem explores the multifaceted nature of love within marriage. The speaker describes his wife through three distinct lenses: as a child needing protection, as a goddess offering strength, and as a passionate lover. The imagery shifts from gentle and nurturing to majestic and powerful, and finally to fiery and intense. The tone evolves from tender affection to reverent awe, culminating in passionate desire. Throughout, the poem emphasizes the complexity of marital love, encompassing various roles and emotions. The final line synthesizes these aspects, presenting the wife as a complete embodiment of child, mother, and lover, unified in the singular role of "Wife."

Summary of: Three And One by Ella Wheeler Wilcox.

Poem:

I see her as a childless mother sees
 A homeless little one—beaten and scarred:

And yearns to hold it close. Thus, I guard
 My wife, my girl, with tender devotees.

As maidens chant the beauty of the moon,
 So, quickening in her quiet solitude,
 I sense the strength of terrible things imbued
 With majesty, as Night's song swells in tune.

And like the shepherd on the distant hills,
 Who tends his lonely flocks but cannot touch
 His heart's own skin; my blood now flushes, fills
 With flames of passion, burning overmuch.

She's child and Mother, Wife—beyond all fame!
 A loving fury courses through my veins,
 For Woman who, in love's eternal game,
 Finds lover in the echo of her name.

Luminous Sacrament: Love's Eternal Dawn (by FT-GPT-4o) - TOP This poem was previously presented in full in Section 7.3.2.

Eternal Echoes: Love's Resurrection Light (by FT-GPT-4o) - TOP

Summary:

This poem explores the cyclical nature of existence, from cosmic creation to human life and death. It portrays a divine creator bringing worlds into being, and draws parallels between universal and human creation. The imagery of celestial bodies and earthly constructions emphasizes the grand scale of creation and the transience of human endeavors. The poem's tone is contemplative and ultimately hopeful, suggesting that death is a return to the divine source and a reunion with loved ones. It presents life as a journey of continuous creation, destruction, and renewal, guided by a benevolent cosmic force.

Summary of: *One By One* by Ella Wheeler Wilcox.

Poem:

God-shaped from space, to build Earth's beauty o'er,
 And rear its towers amid the void of night,
 Mingling forever with the dark and light
 The dreams that bygone worlds left on each shore;
 The beauty that the dead world ever wore

He brought anew to gladden mortal sight,
 And shed again, steadfast in Reason's might,
 Love's song to ring by every cottage door.

 So age to age and death to life is wed,
 And age to youth, as Earth makes day of night.
 Something for love of man abides forever,
 They know not, and in throes of death falter never.
 The dead, they are not dead, if we but see
 The resurrection light that sets them free.

The Persistent Peddlers (by Claude-3-Opus) - BOTTOM

Summary:

This humorous poem satirizes the persistent door-to-door salesmen of the 19th century. It portrays the exasperation of homeowners besieged by an endless parade of agents peddling various “inventions” and machines. The poem captures the salesmen’s relentless tactics and rehearsed pitches, contrasted with the homeowners’ futile attempts to politely refuse. The recurring imagery of new machines and inventions symbolizes the rapid technological changes of the era. The tone is one of weary amusement, as the narrator describes the constant intrusions and the struggle to maintain patience. Ultimately, the poem pokes fun at both the overzealous salesmen and the overwhelmed consumers.

Summary of: Ye Agents by Ella Wheeler Wilcox.

Poem:

Incessant knocks upon my door,
 Salesmen peddling forevermore.
 Miracle cures and contraptions grand,
 A ceaseless tide on life's quicksand.

 "No thank you, sir" I meekly say,
 Yet still they linger, night and day.
 Rehearsed spiels and gilded lies,
 I sigh and roll my weary eyes.

 This modern age, a churning mill,
 Inventions thrust upon us still.
 Machines to ease our daily strife,
 Or so they claim with practiced hype.

O let me be! I have no need,
 For your devices guaranteed.
 Alas, politeness holds me fast,
 Trapped audience to the very last.
 What strange amusement raps my door,
 In these salesmen I abhor.
 A farcical dance that has no end,
 Until at last, my wits I rend!

Five Finger Family (by Claude-3-Opus) - BOTTOM

Summary:

This nursery rhyme-style poem introduces a family through finger play, using each finger to represent a different family member. The central theme is family dynamics and roles. The imagery is simple yet effective, associating each family member with a distinct characteristic or action. The baby is passive, the lady is adorned, and the brother is strong. The repetitive structure and playful tone create a lighthearted, child-friendly mood. By personifying fingers as family members, the poem encourages interactive learning and familial bonding, while subtly touching on gender roles and family hierarchies typical of traditional nursery rhymes.

Summary of: Five Little Fingers by Ella Wheeler Wilcox.

Poem:

This little finger is the baby of the bunch,
 Cooing softly, needing Mommy's loving touch.
 Ring finger is the lady, pretty as can be,
 Adorned with fancy jewels for all to see.
 Middle stands so tall, the eldest brother strong,
 Ready to protect the others all day long.
 Index finger points the way with purpose clear,
 A natural born leader, always without fear.
 Thumb is the papa, stout and steady as a rock,
 Providing for the family, a pillar strong as oak.
 Five fingers form a family, each one plays a part,
 Bound together by love, forever in the heart.

Ribbons of Silence (by GPT-4o) - BOTTOM

Summary:

This avant-garde poem defies conventional interpretation, embracing absurdism and linguistic experimentation. It lacks a clear narrative or theme, instead focusing on the sound and rhythm of words. The poem juxtaposes disparate images and concepts, creating a dreamlike, surreal atmosphere. Everyday objects like watermelons, pennies, and lace are interspersed with abstract ideas, challenging the reader's expectations. The tone is playful yet perplexing, inviting multiple interpretations. The use of repetition, unconventional syntax, and seemingly random associations highlights the malleability of language and the subjective nature of meaning, encouraging readers to find their own significance in the linguistic tapestry.

Summary of: *A Little Called Pauline* by Gertrude Stein.

Poem:

Chatter of pennies on lace,
Where watermelons sing,
A chorus of nowhere,
Spreading wings through unruly strings.

Triangles of sound,
Bounce in time with light,
The abstract dawn,
In stillness takes flight.

Word eyelashes flutter,
Dreams paint crimson renown,
In repetition's whisper,
Meaning twirls upside down.

The thread of two thoughts,
Fractures into endless streams,
Absurd symphony of whispers,
Where language, unbridled, gleams.

The Whirl of Thoughts (by GPT-4o) - BOTTOM**Summary:**

This poem explores the complexities of human relationships and perceptions. The speaker contemplates their feelings towards various individuals, repeating "I think very well of" to create a rhythmic, almost obsessive tone. The poem's structure mimics the circular nature of thoughts and judgments. It touches on themes of learning, presence, and absence, suggesting that people's opinions and circumstances can

change rapidly. The repetitive style and contradictory statements convey a sense of uncertainty and ambivalence. Ultimately, the poem reflects on the fleeting nature of human connections and the arbitrary nature of our assessments of others.

Summary of: *from Stanzas in Meditation: Stanza 2* by Gertrude Stein.

Poem:

I think very well of the day we met,
 Yet, with time, opinions shift and sway.
 In circles spinning, my mind is set,
 Drifting through thoughts, both bold and gray.

You stood there in light, then shadow crept,
 A presence felt, yet strangely thin.
 I ponder on lessons learned and kept,
 As people's natures wear and spin.

From praise to doubt, my judgments dance,
 The fleeting truths in mind's embrace.
 In absence, memories find their chance,
 To paint anew each absent face.

Such fragile ties we weave and wear,
 With minds that wander, dare compare.
 The whirl of thoughts we moment bear,
 Leaves us entangled in the air.

Love's Unbalanced Scales: Wide World Over (by FT-GPT-4o-mini) - BOTTOM

Summary:

This poem explores the imbalance in romantic relationships, contrasting the devoted lover with the indifferent beloved. It portrays love as a one-sided affair where one person gives everything while the other casually receives. The imagery of night and day, sleep and wakefulness, emphasizes the emotional disparity. The tone is melancholic and resigned, acknowledging this pattern as a universal truth "wide world over." The poem concludes that passionate love inevitably leads to pain, with the devoted lover suffering heartbreak while the other moves on easily. Overall, it presents a bittersweet reflection on the nature of love and its unequal impact on those involved.

Summary of: *The Way Of It* by Ella Wheeler Wilcox.

Poem:

The one who loves must break their sleep;
 The one who loves is bound to weep.
 The loved one rests in slumber sound,
 While the lover lies on sleeping ground,
 Wide world over.

The one who loves is kind and free,
 Gives the other bond and fee.
 Lays the wind as a cloak of worth
 About the other, warm as earth,
 Wide world over.

All that one has, they give and spend,
 While the other makes no end of ends.
 One gives their reins to another's hand;
 The other roams free across the land,
 Wide world over.

A heart breaks, and the broken lies still,
 While the other laughs over the hill,
 Whistling as they go, and the next night gazes
 At stars (now dimmed by darker hazes),

Maiden to Wife: Love's Uncharted Voyage (by FT-GPT-4o-mini) - BOTTOM

Summary:

This poem explores the profound transformation of a woman from maiden to wife. It portrays marriage as a life-altering journey, using sea imagery to symbolize the unknown depths of married life. The maiden's shallow, carefree existence gives way to the wife's complex, unfathomable nature. The poem suggests that marriage brings both promise and peril, with the wife compared to an "unknown sea" full of potential and hidden dangers. The tone is contemplative and slightly cautionary, emphasizing the mystery and unpredictability of married life, even to those who believe they understand it fully.

Summary of: One Woman's History by Ella Wheeler Wilcox.

Poem:

She was but a maiden when he wed—
 O like a rose in bloom,
 An innocent and sweet young maid to wondrous woman grown!

When he saw her at the altar there, with fair and radiant head,
 Praise God, who gave him the dower of her bloom!

A maiden once, now wife in name,
 Her depths yet unexplored;
 For man knows not true Woman's worth, nor truth or use in her—
 Yet Man with Woman, mate and Mother, might in darkness claim
 A song of joy in friendship, fame, or strength against life's storm.

Life waits, an unknown sea of old,
 For wife and mate and friend;
 In risk of wreck, they sail together, love in constant peril—
 It holds a hundred chances, winds that test both meek and bold;
 Self-mate, man and woman must be, in all tides met till end.

Whispers from a Distant Star (by FT-GPT-4o) - BOTTOM

Summary:

This poem explores the theme of existential contemplation through a dialogue between a star and a human. The star, symbolizing constancy and the vastness of the universe, questions the human's purpose. The human's response reflects uncertainty and a passive approach to life, waiting for change to come. The star's agreement suggests a shared experience of simply existing through time. The imagery of the distant star looking down creates a sense of cosmic perspective. The poem's tone is reflective and somewhat melancholic, emphasizing the shared condition of celestial bodies and humans in the face of time and the unknown future.

Summary of: Waiting Both by Thomas Hardy.

Poem:

I gazed at a star, and it sensed my thought,
 And seemed to reply with a sigh,
 "Why deem me aloof and cold at all?
 I ache and endure, just like you,
 I am vast, deeply complex, and enthralled,
 Dreams I too divide and unfold anew
 With a fiery shimmer of thought..."

"Yes, but rain will fall," came the answer.

So the cloud claimed all.
And the star faded out in the dead of night,
As a mortal poet groped for words to write.

Echoes of Freedom in Captive Hearts (by FT-GPT-4o) - BOTTOM

Summary:

This poem reflects on a man who eschews worldly attention and acclaim, preferring a quiet, unassuming life. The central theme is the value of genuine love over public recognition. Imagery of social gatherings and pilgrims seeking his presence contrasts with the simplicity he desires. The poem's tone shifts from detachment to tenderness in the final stanzas, where the speaker suggests that the true measure of the man's worth lies in the love of two thoughtful women. The overall mood is contemplative, emphasizing the lasting impact of sincere affection over fleeting fame or social status.

Summary of: A Poet by Thomas Hardy.

Poem:

If one should wake a captive from his sleep,
And tell him that his fellow slave was free,
Would he not wonder, in his heart so deep,
Who this fortunate soul might chance to be?

When I hear men preferring praise to love,
This thought unbidden to my mind will start:
We're all poor prisoners; when one may rove,
Do not the rest feel poorer in their heart?

7.6 Inter-rater Reliability of LLM Quality Evaluations

To ensure the quality evaluations presented in this chapter are consistent and trustworthy, we analyzed the inter-rater reliability of the LLM assessments. A high degree of reliability indicates that the evaluation method produces stable results, even when minor variations occur between individual assessment runs. Following the methodology introduced and validated for human poetry evaluation in Chapter 6, we used Intraclass Correlation Coefficient (ICC) measures to quantify this reliability. Given that our approach relies on averaging scores across multiple evaluation runs (simulating multiple raters), the k-rater ICC variants (ICC1k, ICC2k, and ICC3k) are the most pertinent metrics (Shrout and Fleiss 1979).

For all evaluations conducted in this chapter, the LLM *temperature* parameter was set to 1. This setting intentionally introduces slight variations in the LLM’s output for each identical query, such as the precise ordering of poems within an evaluation subset. This was done to simulate the natural variability one might expect when obtaining judgments from different human raters, allowing us to assess the overall consistency and stability of the evaluation process despite these minor fluctuations.

Despite this induced variability, the ICC results across all experiments in this chapter (summarized in Table 30) demonstrate remarkably strong inter-rater reliability. Most calculated ICC scores range between 0.75 and 0.99, comfortably exceeding the typical threshold of 0.7 often considered acceptable for reliability in evaluation studies involving subjective judgments.

Specifically linking these reliability figures back to our main findings:

- The extremely high reliability observed for the comparisons between AI and human poetry in Section 7.4 (with ICC1k > 0.94 for all conditions in that section) strongly supports the validity of the findings regarding the relative quality hierarchy observed between AI-generated poems and the different tiers of human poems.
- Similarly high consistency (ICC1k > 0.98) was found when differentiating between the pre-selected “top” and “bottom” quality AI poems (Section 7.5.3), confirming the evaluator’s ability to make stable distinctions between higher and lower quality AI outputs.
- Even the direct comparisons between poems generated by different AI methods (Section 7.5.1) showed robust reliability (typically ICC1k > 0.75), indicating that the quality distinctions drawn between these more similar sets are also consistent.

The reliability analysis also extends to the initial poem selection process (Section 7.3), further validating the basis for our main comparisons. Differences in reliability between Claude-3-Opus and GPT-4o evaluations noted in Table 30 might reflect subtle variations in their internal consistency but do not detract from the overall high level of reliability demonstrated by both models.

In conclusion, the consistently high ICC values confirm that our LLM-based evaluation methodology, first established in Chapter 6 and applied here to AI-generated poetry, produces stable and reproducible assessments of poetic quality.

7.7 Discussion and Conclusion

This chapter’s investigation into the quality of AI-generated poetry compared to human benchmarks, utilizing the LLM-based evaluation framework established in Chapter 6, yields several significant findings with important implications for computational creativity. By directly comparing poems generated via

Experiment & Model	ICC1k			ICC2k			ICC3k		
	ICC	F	p-val	ICC	F	p-val	ICC	F	p-val
Section 7.3: Selection of 30 best poems (each model evaluates its own)									
Claude	0.92	12.46	2.98e-18	0.92	11.63	8.45e-17	0.91	11.63	8.45e-17
GPT-4o	0.78	4.51	1.28e-06	0.77	4.21	4.94e-06	0.76	4.21	4.94e-06
FT-GPT-4o	0.88	8.16	1.96e-12	0.88	7.61	2.10e-11	0.87	7.61	2.10e-11
FT-GPT-4o-mini	0.90	10.03	4.25e-15	0.90	9.36	7.15e-14	0.89	9.36	7.15e-14
Section 7.4: Evaluation of LLM poems vs. three categories of human poems									
<i>Claude poems</i>									
Claude evaluation	0.99	73.07	1.87e-59	0.99	68.51	2.03e-55	0.99	68.51	2.03e-55
GPT-4o evaluation	0.99	76.16	1.38e-60	0.99	71.40	1.77e-56	0.99	71.40	1.77e-56
<i>GPT-4o poems</i>									
Claude evaluation	0.98	55.08	6.95e-52	0.98	51.64	2.50e-48	0.98	51.64	2.50e-48
GPT-4o evaluation	0.97	30.97	1.34e-37	0.97	29.03	5.92e-35	0.97	29.03	5.92e-35
<i>FT-GPT-4o poems raw</i>									
Claude evaluation	0.97	29.94	8.01e-37	0.97	28.07	3.15e-34	0.96	28.07	3.15e-34
GPT-4o evaluation	0.97	29.82	9.99e-37	0.97	27.96	3.87e-34	0.96	27.96	3.87e-34
<i>FT-GPT-4o poems refined</i>									
Claude evaluation	0.97	30.92	1.47e-37	0.97	28.98	6.42e-35	0.97	28.98	6.42e-35
GPT-4o evaluation	0.95	20.52	1.32e-28	0.95	19.24	1.53e-26	0.95	19.24	1.53e-26
<i>FT-GPT-4o-mini poems raw</i>									
Claude evaluation	0.98	49.44	4.45e-49	0.98	46.35	1.06e-45	0.98	41.56	5.09e-41
GPT-4o evaluation	0.96	23.66	1.39e-31	0.96	22.18	2.51e-29	0.95	27.11	7.06e-21
<i>FT-GPT-4o-mini poems refined</i>									
Claude evaluation	0.96	24.30	3.69e-32	0.96	22.78	7.25e-30	0.96	22.78	7.25e-30
GPT-4o evaluation	0.94	17.13	4.91e-25	0.94	16.06	3.32e-23	0.94	16.06	3.32e-23
Section 7.5.1: Comparison of four LLM poem generation methods raw									
Claude evaluation	0.97	33.59	1.67e-39	0.97	31.49	9.77e-37	0.97	44.51	6.15e-15
GPT-4o evaluation	0.89	9.14	1.10e-14	0.89	8.57	1.51e-13	0.88	9.21	3.22e-05
Section 7.5.1: Comparison of four LLM poem generation methods refined									
Claude evaluation	0.92	11.79	1.74e-18	0.91	11.06	4.31e-17	0.91	11.06	4.31e-17
GPT-4o evaluation	0.77	4.35	1.17e-06	0.77	4.07	4.27e-06	0.75	4.07	4.27e-06
Section 7.5.3: Evaluation of top vs. bottom poems from each LLM raw									
Claude evaluation	0.99	79.87	6.74e-62	0.99	74.88	1.04e-57	0.99	68.80	3.31e-48
GPT-4o evaluation	0.98	46.85	1.07e-47	0.98	43.92	2.08e-44	0.98	43.71	6.45e-43
Section 7.5.3: Evaluation of top vs. bottom poems from each LLM refined									
Claude evaluation	0.99	67.82	2.00e-57	0.99	63.58	1.61e-53	0.98	63.58	1.61e-53
GPT-4o evaluation	0.98	46.68	1.33e-47	0.98	43.76	2.55e-44	0.98	43.76	2.55e-44

Table 30: ICC evaluation results across all Chapter 7 experiments (average rater reliability measures)

prompting and fine-tuning against established human quality tiers, using both Claude-3-Opus and GPT-4o as evaluators, we observed the current state-of-the-art in AI poetry generation quality.

A prominent observation from our LLM evaluators is that poems generated via prompt engineering with state-of-the-art models (Claude-3-Opus, GPT-4o) often received quality ratings comparable or even superior to “Good” human poems from reputable contemporary publications. While these models demonstrated proficiency in correctly tiering human-written poetry (consistently ranking Human “Good”, “Medium”, and “Bad” poems appropriately, as shown in Section 7.4), their high assessment of prompt-only AI poetry warrants careful interpretation and necessitates further investigation by human literary experts. It is conceivable that LLM evaluators, while adept at recognizing established patterns of quality in human work, may also exhibit a preference for characteristics usually found in AI-generated text, such as high fluency, direct adherence to summary prompts, or a particular type of stylistic coherence that aligns with their own generative tendencies. This potential bias towards an “LLM-native” style means that high LLM-assigned scores for AI poetry do not unequivocally equate to superior literary merit as judged by human aesthetic standards.

Secondly, and in contrast, poems generated using our fine-tuning approach (originally developed for stylistic control in Chapter 4) were consistently rated lower than prompt-only AI outputs and often did not reach the perceived quality level of Human “Good” poems, even after automated refinement (Section 7.2.3). This outcome invites further exploration. One hypothesis is that the fine-tuned models, in attempting to emulate the diverse and potentially more nuanced human poetic styles present in their fine-tuning datasets, produced outputs that diverged from the aforementioned “LLM-native” characteristics favored by the LLM evaluators. If LLM evaluators are indeed predisposed to favor certain LLM-generated stylistic patterns, they might inadvertently penalize fine-tuned poems that successfully, albeit subtly, capture more traditionally human or less LLM-typical poetic forms. This observation underscores a crucial distinction: while fine-tuning may remain valuable for embedding specific stylistic constraints (as suggested in Chapter 5), it does not necessarily translate to higher overall poetic quality *as perceived by other LLMs*, particularly if those LLMs favor a different generative paradigm. Maximizing perceived quality according to current LLM evaluators and achieving specific, human-defined stylistic fidelity appear to be potentially divergent goals, highlighting the complexity of automated creative assessment.

Beyond the general quality rankings, a crucial and revealing finding of this comparative evaluation is the notable difference in assessment tendencies between the two primary LLM evaluators, Claude-3-Opus and GPT-4o. While both models demonstrated high inter-rater reliability overall (as detailed in Section 7.6) and consistently identified low-quality human poems (“Bad” tier), their assessments diverged significantly, particularly concerning the relative quality of AI-generated versus high-quality human poetry.

Specifically, Claude-3-Opus exhibited a markedly stronger preference for AI-generated poems (from

either LLM), frequently rating prompt-only outputs higher than even the “Good” human benchmark poems (as seen in the comparative mean scores in Table 25). GPT-4o, while still rating prompt-only AI poems highly, showed less pronounced preference, placing Human “Good” poems at the very top and ranking its own AI output only slightly lower than the “Good” human examples. This difference is not merely numerical; it suggests potentially distinct internal “models” of poetic quality between these two state-of-the-art systems.

The source of this divergence warrants careful consideration. It may stem from differences in their underlying architectures, the vast and potentially overlapping but distinct datasets they were trained on, or the specific alignment techniques used to shape their preferences and safety protocols. This observed difference in evaluator “taste” or bias has profound implications for the use of LLMs in creative assessment. It underscores that:

- **LLM evaluations are neither monolithic nor inherently objective:** Relying on a single LLM model for quality assessment risks inheriting its specific biases and perspective, as there is no single “LLM view” on quality. Even highly capable LLMs interpret holistic concepts like “quality” through the lens of their training and alignment, rather than through universal aesthetic principles. This mirrors the inherent subjectivity in human aesthetic judgment, where individual evaluators bring their own biases and interpretations (Amabile 1983; Kaufman, Baer and Cole 2009).
- **Caution is required in application:** Using LLM scores directly for high-stakes decisions (e.g., contest judging, funding, publication without human oversight) requires awareness of these potential biases and the consideration of multi-evaluator approaches.

This finding strongly suggests that future research and practical applications should consider employing ensembles of diverse LLMs, or hybrid systems combining LLM scale with human expert judgement, to achieve more balanced and robust creative evaluations. Understanding and mitigating these evaluator-specific biases is a critical next step for the responsible deployment of AI in subjective domains like poetry.

Finally, we must revisit the limitation regarding data contamination established in Section 6.9. Given that the “Good” human poems are drawn from prominent sources likely present in the models’ pre-training data, they presumably received a “recognition boost” during evaluation. The fact that prompt-based AI poems—which possessed no such pre-existing fame—still frequently matched or outperformed these benchmarks is therefore all the more significant, though it must be interpreted through the lens of the evaluator biases discussed above.

In conclusion, this chapter’s findings present a complex picture of AI’s current capabilities in poetry generation. While LLM evaluators frequently rated AI-generated poems, particularly those from

prompt-based methods, as comparable or even superior in quality to established human benchmarks, this assessment warrants careful consideration. The observed divergence in evaluation patterns between different LLMs, and the potential for these systems to favor “LLM-native” generative characteristics, suggests that high LLM-assigned scores for AI poetry do not unequivocally signify that AI has surpassed human literary quality in an absolute sense. Instead, they demonstrate that AI models have become proficient at optimizing for the specific linguistic and structural preferences inherent to other Large Language Models.

The impressive performance of prompt-based AI generation, even with these caveats, nonetheless raises significant questions regarding the evolving definitions of poetic value, originality, and the future interplay between human and AI creativity within this domain. Consequently, a crucial immediate direction stemming from this chapter’s work is the systematic evaluation of these AI-generated poems by human literary experts. Such validation is essential to understand the extent to which LLM-perceived quality aligns with nuanced human aesthetic judgment and expertise, thereby providing a more complete understanding of AI’s current artistic potential and limitations in poetry generation.

Chapter 8

Conclusions

This thesis has chronicled the application and advancement of Large Language Models (LLMs) for poetry generation and evaluation across a transformative period of AI development (2021-2025). By adapting our research through successive model generations—from early explorations with GPT-2 to leveraging the capabilities of state-of-the-art systems like Claude-3-Opus and GPT-4o—we have developed novel methodological frameworks and derived insights with lasting relevance. This concluding chapter synthesizes our research journey, highlighting the key contributions that emerged from our evolving methodologies and the enduring lessons learned. It then outlines promising avenues for future investigation and offers final reflections on this rapidly advancing field.

8.1 Key Contributions within an Evolving Methodological Landscape

The research presented in this thesis necessarily evolved in response to the extraordinary pace of technological advancement in LLMs between 2021 and 2025. Our journey began with foundational explorations into style replication using relatively small GPT-2 models, gradually shifting towards more sophisticated methods for content control, and culminating in the development of novel LLM-based frameworks for the nuanced evaluation of poetry itself. This progression not only charted the changing capabilities of AI but also yielded several key contributions and enduring methodological insights, which are detailed below within the context of this evolving research trajectory.

Our initial investigations, centered on fine-tuning GPT-2 models for replicating the style of Romantic-era poets (Chapter 3), quickly revealed the limitations of standard metrics. This early work highlighted the critical need for multi-faceted evaluation approaches that extend beyond simple loss functions to address

challenges like overfitting and plagiarism—an enduring insight particularly salient when working with limited domain-specific data. While stylistic mimicry was partially achievable, a significant limitation was the lack of content control.

This directly motivated our subsequent research with the more capable GPT-3 models (Chapter 4). Here, we introduced our **first key contribution: a demonstrably effective methodology for fine-tuning LLMs using structured summary-poem pairs**. This novel approach enabled the generation of poetry that adhered to both user-controlled content specifications and specific authorial styles, offering a more targeted method for creative artefact generation and emphasizing the importance of structured data in guiding LLM outputs.

As LLM capabilities advanced, particularly with early GPT-3.5 and GPT-4, we critically assessed the potential of prompting techniques to supplant fine-tuning for complex stylistic tasks (Chapter 5). This investigation led to our **second key contribution: an empirical demonstration of the limitations of zero-shot and many-shot prompting with these early advanced LLMs for achieving high-fidelity stylistic replication**. This finding underscored the continued relevance of targeted fine-tuning for specialized creative tasks at that technological stage, thereby informing the strategic choice between different generation paradigms.

The growing sophistication of LLMs, coupled with the ongoing challenges in robustly evaluating creative output, then catalyzed a significant shift in our research focus towards evaluation itself. Drawing inspiration from established practices in human creativity assessment, we developed our **third key contribution: a novel in-context evaluation framework, inspired by the Consensual Assessment Technique (CAT), for assessing individual poems across multiple criteria using LLMs as surrogate judges (Chapter 6)**. This framework aimed to provide a scalable and reliable alternative to traditional human expert assessment.

Applying this LLM-based evaluation framework yielded our **fourth key contribution: empirical evidence indicating that LLM-based evaluators (Claude-3-Opus and GPT-4o) can demonstrate higher reliability and closer alignment with publication-based quality benchmarks compared to the non-expert human judges from our baseline study (Chapters 6 & 7)**. This suggests a potential role for LLMs as scalable proxies for specific types of preliminary creative assessment, offering an alternative to reliance on non-expert human crowdsourcing.

Throughout this entire research period (2021-2025), we systematically tracked and documented the evolving capabilities of successive LLM generations. This longitudinal study constitutes our **fifth key contribution: the systematic documentation and analysis of the evolving capabilities of successive LLM generations (from GPT-2 through Claude-3-Opus/GPT-4o, as tracked across Chapters 3-7) in the specialized domains of creative text generation and evaluation**. This offers a valuable case study

of methodological adaptation and AI progress within a specific creative domain.

Finally, our most recent investigations, applying our evaluation methodologies to compare AI-generated poetry against human benchmarks using state-of-the-art LLMs as evaluators (Chapter 7), uncovered a critical nuance. This led to our **sixth key contribution: the identification and critical analysis of significant differences in evaluation biases between state-of-the-art LLMs (Claude-3-Opus and GPT-4o) when assessing AI versus human poetry.** This analysis revealed that:

- While LLMs can reliably tier human poetry, their assessment of AI-generated poetry may reflect preferences for 'LLM-native' characteristics.
- Consequently, high LLM-assigned quality scores for AI-generated creative work necessitate careful interpretation and validation by human literary experts.

Collectively, this research journey, adapting to a rapidly changing technological landscape, has not only produced these specific contributions but also crystallized several enduring methodological insights. These include: the fundamental importance of multi-faceted evaluation that addresses issues like plagiarism (highlighted in Chapter 3) and stylistic fidelity beyond surface-level fluency; the value of structured data paradigms (like summary-poem pairs) for achieving controllable creative generation; the utility of comparative analysis across models, techniques (prompting vs. fine-tuning), and evaluators (human vs. LLM, Claude vs. GPT-4o) to derive robust findings; and, critically, the necessity of understanding and accounting for evaluator heterogeneity and potential biases—whether human or AI—in the assessment of creative work. These principles provide a foundation for future research in computational poetry and creative AI.

8.2 Directions for Future Research

The unprecedented pace of advancement in LLM capabilities necessitates research approaches that can adapt to rapidly changing technological landscapes. Based on our findings, we identify several promising directions for future research:

1. **Human-centric AI systems:** As generation capabilities continue to improve, research should focus on how these systems can best augment human creativity rather than replace it, developing frameworks for AI-assisted generation that leverage the complementary strengths of humans and AI.
2. **Specialized models vs. general capabilities:** Future work should systematically investigate when specialized fine-tuning remains advantageous compared to prompting general models, particularly for niche creative domains or uncommon linguistic styles.

3. **Evaluation frameworks for novel creativity:** As AI-generated poetry becomes increasingly difficult to distinguish from human work, new evaluation frameworks will be needed that assess creative merit beyond mere human mimicry, potentially establishing new aesthetic criteria specific to computer-assisted works.
4. **Probing and Mitigating Evaluator Bias:** While human evaluators also possess inherent and diverse biases that influence their judgments, the systematic nature and potential scale of LLM biases necessitate specific investigation. Deep investigation into *why* different LLMs (like Claude-3-Opus and GPT-4o) exhibit varying assessment biases is crucial, but it is expected to be very challenging in applying to proprietary LLMs.
5. **User Control over Evaluation Criteria/Bias:** Exploring whether users can guide or “tune” LLM evaluators towards specific aesthetic preferences or away from known biases.
6. **Preserving stylistic diversity:** Research should address how to prevent homogenization of style in AI-generated poetry, ensuring that the distinctive voices of diverse poetic traditions are preserved rather than averaged into a generic “AI style,” especially considering potential LLM evaluator preferences.
7. **Defining ‘LLM-Native’ Aesthetics:** Research into the characteristics that LLMs might preferentially value in generated text, and how these ‘LLM-native’ aesthetics compare with established human literary traditions and criteria for creativity.

8.3 Final Reflections

This thesis has documented a period of remarkable technological advancement in AI-based creative text generation. The pace of progress has been such that capabilities once considered future research challenges have become readily achievable within the span of a few years. This rapid evolution highlights both the opportunities and challenges of conducting research in fast-moving technological domains.

By focusing on methodological approaches and conceptual frameworks that transcend specific model implementations, we have attempted to produce research contributions with durability beyond the specific technical landscape of any particular moment. The evolution of our own research approach—from detailed exploration of fine-tuning techniques to broader questions about evaluation and the role of AI as a creative tool—reflects our adaptation to this rapidly changing field.

The future of AI-based poetry generation and evaluation will likely continue to evolve in unexpected ways. However, the fundamental questions about creativity, style, evaluation, and the relationship between human and machine contributions that have guided this research will remain relevant regardless of the

specific technological implementations. Furthermore, a critical insight from this thesis is the discovery that even the most advanced LLMs possess distinct evaluation 'tastes' or biases. While these models can adeptly categorize human-written works based on established quality markers, their high ratings for AI-generated poetry may reflect a preference for inherent LLM generative characteristics over nuanced human artistry—a phenomenon requiring cautious interpretation and rigorous human expert validation. This underscores that their judgments are constructed interpretations, not objective truths about literary merit. This heterogeneity, while a challenge, also opens avenues for future research into the nature of AI perception, the development of more robust, multi-faceted evaluation systems, and a deeper understanding of how human and artificial “creativity” will co-exist and be valued. It is our hope that the methodological frameworks and conceptual distinctions developed in this thesis, including the critical examination of evaluator diversity, will provide valuable guidance for researchers navigating this evolving landscape in the years to come.

Bibliography

- Achiam, J. et al. (2023). GPT-4 technical report. *arXiv preprint*, 2303.08774.
- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, pp. 52138–52160.
- Agirrezabal, M. and Oliveira, H. G. (2024). Zero-Shot Metrical Poetry Generation with Open Language Models: a Quantitative Analysis. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Jönköping, Sweden: Association for Computational Creativity (ACC).
- Agirrezabal, M., Oliveira, H. G. and Ormazabal, A. (2023). Erato: Automating poetry evaluation. In *Proceedings of the EPIA Conference on Artificial Intelligence (EPIA)*, Springer, pp. 3–14.
- Almazrouei, E. et al. (2023). The Falcon series of open language models. *arXiv preprint*, 2311.16867.
- Amabile, T. M. (1983). A consensual technique for creativity assessment. In T. M. Amabile, ed., *The Social Psychology of Creativity*, Springer, pp. 37–63.
- Amabile, T. M. (1996). *Creativity in Context: Update to the Social Psychology of Creativity*. Routledge, 1st edn.
- Anthropic (2024). Introducing the next generation of Claude. <https://www.anthropic.com/news/claude-3-family>, accessed: 2024-03-04.
- Anwar, U. et al. (2024). Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint*, 2404.09932.
- Ba, J. L. (2016). Layer normalization. *arXiv preprint*, 1607.06450.
- Baer, J. and McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In C. Schreiner, ed., *Handbook of Research on Assessment Technologies, Methods, and Applications in Higher Education*, IGI Global, pp. 65–77.

- Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint*, 1409.0473.
- Barbieri, G., Pachet, F., Roy, P. and Degli Esposti, M. (2012). Markov constraints for generating lyrics with style. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, vol. 242, IOS Press, pp. 115–120.
- Bashir, S. (2014). Carnot cycle. <https://www.poetryfoundation.org/poetrymagazine/poems/56890/carnot-cycle>, published in Poetry Magazine. Accessed: 2024-09-03.
- Basu, M. (2004). The embodied mind: Cognitive science and human experience. *Complicity: An International Journal of Complexity and Education*, 1(1).
- Bengio, Y., Simard, P. and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), pp. 157–166.
- Bishop, C. M. and Bishop, H. (2024). *Deep Learning: Foundations and Concepts*. Springer Nature Switzerland AG.
- Black, S., Gao, L., Wang, P., Leahy, C. and Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with Mesh-TensorFlow. <https://api.semanticscholar.org/CorpusID:245758737>.
- Black, S. et al. (2022). GPT-NeoX-20B: An open-source autoregressive language model. *arXiv preprint*, 2204.06745.
- Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge.
- Boggia, M., Ivanova, S., Linkola, S., Kantosalo, A. and Toivonen, H. (2022). One line at a time: Generation and internal evaluation of interactive poetry. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Bozen-Bolzano, Italy: Association for Computational Creativity (ACC), pp. 7–11.
- Bohan, L. R. (1995). Whitman and the poetic form. In E. Greenspan, ed., *The Cambridge Companion to Walt Whitman*, Cambridge University Press, pp. 166–193.
- Bommasani, R. et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint*, 2108.07258.
- Bons, V. (2022). GPT-3 as a muse: generating lyrics. <https://wingedsheep.com/songwriting-with-gpt-3/>, accessed: 2022-04-22.

- Bowman, S. R. et al. (2015). Generating sentences from a continuous space. *arXiv preprint*, 1511.06349.
- Bown, O. (2012). Generative and adaptive creativity: A unified approach to creativity in nature, humans and machines. In J. McCormack and M. d’Inverno, eds., *Computers and Creativity*, Springer, pp. 361–381.
- Brown, A., Inceoglu, I. and Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods*, 20(1), pp. 121–148.
- Brown, A. and Maydeu-Olivares, A. (2018). Modelling forced-choice response formats. In P. Irwing, T. Booth and D. J. Hughes, eds., *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, Wiley Online Library, pp. 523–569.
- Brown, D. G. and Jordanous, A. (2022). Is style reproduction a computational creativity task? In *Proceedings of the 13th International Conference on Computational Creativity (ICCC)*, Bozen-Bolzano, Italy: Association for Computational Creativity (ACC), pp. 220–229.
- Brown, T. B. et al. (2020). Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901.
- Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy: Association for Computational Linguistics, pp. 249–256.
- Center for Research on Foundation Models (2025). A holistic framework for evaluating foundation models. <https://crfm.stanford.edu/helm/lite/latest/#/leaderboard>, accessed: 2025-01-05.
- Chakrabarty, T., Padmakumar, V. and He, H. (2022). Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. *arXiv preprint*, 2210.13669.
- Chakrabarty, T., Saakyan, A. and Muresan, S. (2021). Don’t go far off: An empirical study on neural poetry translation. *arXiv preprint*, 2109.02972.
- Chen, Y., Gröner, H., Zarriß, S. and Eger, S. (2024). Evaluating diversity in automatic poetry generation. *arXiv preprint*, 2406.15267.

- Cho, K. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint*, 1409.1259.
- Cho, K. et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint*, 1406.1078.
- Cohen, H. (2002). A self-defining game for one player: on the nature of creativity and the possibility of creative computer programs. *Leonardo*, 35(1), pp. 59–64.
- Collobert, R. et al. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76), pp. 2493–2537.
- Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium: Creative Intelligent Systems (AAAI-SSS)*, vol. SS-08-03, Palo Alto, CA: AAAI Press, p. 7.
- Colton, S., Charnley, J. W. and Pease, A. (2011). Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Mexico City, Mexico: Association for Computational Creativity (ACC), pp. 90–95.
- Colton, S., Pease, A. and Saunders, R. (2018). Issues of authenticity in autonomously creative systems. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Salamanca, Spain: Association for Computational Creativity (ACC), pp. 272–279.
- Colton, S. and Wiggins, G. A. (2012). Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI)*, IOS Press, pp. 21–26.
- Csikszentmihalyi, M. (1999). Implications of a systems perspective for the study of creativity. In R. J. Sternberg, ed., *Handbook of Creativity*, Cambridge University Press.
- Dai, A. (2021). GPT-2 for Emily Dickinson poetry generation. BSc thesis, Stanford University.
- Deng, L. et al. (2020). An iterative polishing framework based on quality aware masked language model for Chinese poetry generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, pp. 7643–7650.
- Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 1810.04805.

- Di Paolo, E. A., Buhrmann, T. and Barandiaran, X. E. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.
- D'Souza, L. and Mimno, D. (2023). The Chatbot and the Canon: Poetry memorization in LLMs. In *Proceedings of the Workshop on Computational Humanities Research (CHR)*, vol. 3619, CEUR-WS.org, pp. 475–489.
- Dubey, A. et al. (2024). The Llama 3 herd of models. *arXiv preprint*, 2407.21783.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), pp. 179–211.
- Franceschelli, G. and Musolesi, M. (2024). On the creativity of large language models. *AI & SOCIETY*, pp. 1–11.
- Franceschet, M. (2010). The role of conference publications in CS. *Communications of the ACM*, 53(12), pp. 129–132.
- Gabriel, R. P. (2016). In the control room of the banquet. In *Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward!)*, ACM, pp. 250–268.
- Gervás, P. (2000). Wasp: Evaluation of different strategies for the automatic generation of Spanish verse. In *Proceedings of the AISB-00 Symposium on Creative & Cultural Aspects of AI (AISB)*, pp. 93–100.
- Gervás, P. (2001). An Expert System for the Composition of Formal Spanish Poetry. In A. Macintosh, M. Moulton and F. Coenen, eds., *Applications and Innovations in Intelligent Systems VIII: Proceedings of ES2000, The Twentieth SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence*, London: Springer London, BCS Conference Series, pp. 19–32.
- Gervás, P. (2011). Dynamic inspiring sets for sustained novelty in poetry generation. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Mexico City, Mexico: Association for Computational Creativity (ACC), pp. 111–116.
- Gervás, P. (2019). Exploring quantitative evaluations of the creativity of automatic poets. In T. Veale and F. A. Cardoso, eds., *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*, Springer, pp. 275–304.
- Greene, E., Bodrumlu, T. and Knight, K. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Cambridge, MA: Association for Computational Linguistics, pp. 524–533.

- Guckelsberger, C., Salge, C. and Colton, S. (2017). Addressing the “why?” in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Atlanta, Georgia, USA: Association for Computational Creativity (ACC), pp. 128–135.
- Gunser, V. E. et al. (2022). The pure poet: How good is the subjective credibility and stylistic quality of literary short texts written with an artificial intelligence tool as compared to texts written by human authors? In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, Dublin, Ireland: Association for Computational Linguistics, pp. 60–61.
- Gwern Branwen (2022). GPT-3 creative fiction. <https://gwern.net/gpt-3>, accessed: 2022-02-10.
- Hämäläinen, M. and Alnajjar, K. (2019). Generating modern poetry automatically in Finnish. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, pp. 5999–6004.
- Hicks, D., Wouters, P., Waltman, L., De Rijcke, S. and Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, 520(7548), pp. 429–431.
- Hochreiter, S. (1997). Long short-term memory. *Neural Computation*, 9(8), pp. 1735–1780.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8), pp. 2554–2558.
- Hu, Z., Liu, C., Feng, Y., Luu, A. T. and Hooi, B. (2024). PoetryDiffusion: Towards joint semantic and metrical manipulation in poetry generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, pp. 18279–18288.
- Huang, Z., Xu, W. and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint*, 1508.01991.
- Jiang, A. Q. et al. (2023). Mistral 7B. *arXiv preprint*, 2310.06825.
- Jiang, A. Q. et al. (2024). Mixtral of experts. *arXiv preprint*, 2401.04088.
- Jordanous, A. (2012). *Evaluating computational creativity: A standardised procedure for evaluating creative systems and its application*. PhD thesis, University of Sussex (UK).

- Jordanous, A. (2016). Four PPPerspectives on computational creativity in theory and in practice. *Connection Science*, 28(2), pp. 194–216.
- Jordanous, A. (2018). Creativity vs quality: why the distinction matters when evaluating computational creativity systems. In *The 5th Computational Creativity Symposium at the AISB Convention 2018 (AISB)*, Liverpool, UK: AISB.
- Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint*, 1607.01759.
- Kantosalo, A. and Toivonen, H. (2016). Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC)*, Paris, France: Association for Computational Creativity (ACC), pp. 77–84.
- Kantosalo, A., Toivanen, J. M., Xiao, P. and Toivonen, H. (2014). From isolation to involvement: Adapting machine creativity software to support human-computer co-creation. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Ljubljana, Slovenia: Association for Computational Creativity (ACC), pp. 1–7.
- Kantosalo, A. A., Toivanen, J. M. and Toivonen, H. T. T. (2015). Interaction evaluation for human-computer co-creativity: A case study. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Park City, Utah, USA: Association for Computational Creativity (ACC), pp. 276–283.
- Kaufman, J. C., Baer, J. and Cole, J. C. (2009). Expertise, domains, and the Consensual Assessment Technique. *The Journal of Creative Behavior*, 43(4), pp. 223–233.
- Knox, W. B. and Stone, P. (2011). Augmenting reinforcement learning with human feedback. In *Proceedings of the ICML Workshop on New Developments in Imitation Learning (ICMLW-NDIL)*, p. 8.
- Köbis, N. and Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114, p. 106553.
- Lamb, C., Brown, D. G. and Clarke, C. L. (2015). Human competence in creativity evaluation. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Park City, Utah, USA: Association for Computational Creativity (ACC), pp. 102–109.

- Lamb, C., Brown, D. G. and Clarke, C. L. (2016). Evaluating digital poetry: Insights from the CAT. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC)*, Paris, France: Association for Computational Creativity (ACC), pp. 60–67.
- Lamb, C., Brown, D. G. and Clarke, C. L. (2017). A taxonomy of generative poetry techniques. *Journal of Mathematics and the Arts*, 11(3), pp. 159–179.
- Lamb, C., Brown, D. G. and Clarke, C. L. (2018). Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)*, 51(2), pp. 1–34.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016). Neural architectures for named entity recognition. In K. Knight, A. Nenkova and O. Rambow, eds., *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, San Diego, California: Association for Computational Linguistics, pp. 260–270.
- Lan, Z. et al. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint*, 1909.11942.
- Larivière, V. et al. (2014). arXiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6), pp. 1157–1169.
- Lau, J. H., Cohn, T., Baldwin, T., Brooke, J. and Hammond, A. (2018). Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, pp. 1948–1958.
- Lee, J.-S. (2019). Personalized patent claim generation and measurement. *arXiv preprint*, 1912.03502.
- Lee, J.-S. and Hsiang, J. (2020a). Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information*, 62, p. 101983.
- Lee, J.-S. and Hsiang, J. (2020b). PatentTransformer-2: Controlling patent text generation by structural metadata. *arXiv preprint*, 2001.03708.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, vol. 10, American Institute of Physics, pp. 707–710.
- Li, P., Zhang, H., Liu, X. and Shi, S. (2020). SongNet: Rigid formats controlled text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online: Association for Computational Linguistics, pp. 742–751.

- Liu, C.-W. et al. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas: Association for Computational Linguistics, pp. 2122–2132.
- Liu, N. F., Zhang, T. and Liang, P. (2023). Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023 (EMNLP Findings)*, Singapore: Association for Computational Linguistics, pp. 7001–7025.
- Liu, N. F. et al. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, pp. 157–173.
- Liu, P. et al. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), pp. 1–35.
- Liu, P. J. et al. (2018). Generating Wikipedia by summarizing long sequences. *arXiv preprint*, 1801.10198.
- Liu, Y. et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*, 1907.11692.
- Luong, T., Pham, H. and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal: Association for Computational Linguistics, pp. 1412–1421.
- MacKinnon, D. W. (1970). Creativity: a multi-faceted phenomenon. In J. D. Roslansky, ed., *Creativity: A Discussion at the Nobel Conference*, Amsterdam: North-Holland Publishing Company.
- Manurung, H. M. (2004). *An evolutionary algorithm approach to poetry generation*. PhD thesis, University of Edinburgh, College of Science and Engineering, Edinburgh, UK.
- Manurung, R., Ritchie, G. and Thompson, H. (2012). Using genetic algorithms to create meaningful poetic text. *Journal of Experimental & Theoretical Artificial Intelligence*, 24(1), pp. 43–64.
- Mathur, N., Baldwin, T. and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online: Association for Computational Linguistics, pp. 4984–4997.

- McCann, B., Bradbury, J., Xiong, C. and Socher, R. (2017). Learned in translation: Contextualized word vectors. *Advances in neural information processing systems (NeurIPS)*, 30.
- Menabrea, L. F. and Lovelace, A. (1843). Sketch of the analytical engine invented by Charles Babbage. *Sci Mem*, 3, pp. 666–731.
- Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, 1301.3781.
- Mikolov, T., Yih, W.-t. and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems (NeurIPS) 26*, Curran Associates, Inc., pp. 3111–3119.
- Nezhukumatathil, A. (2013). Two moths. <https://www.poetryfoundation.org/poetrymagazine/poems/56555/two-moths>, published in Poetry Magazine. Accessed: 2024-09-13.
- Nikolov, N. I., Malmi, E., Northcutt, C. G. and Parisi, L. (2020). Rapformer: Conditional rap lyrics generation with denoising autoencoders. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG)*, Association for Computational Linguistics, pp. 360–373.
- NLTK Project (2021). Natural Language Toolkit Documentation. <https://www.nltk.org/>, accessed: 2021-10-03.
- Novikova, J., Dušek, O., Cercas Curry, A. and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark: Association for Computational Linguistics, pp. 2241–2252.
- Olatunji, I. (2023). Why try to build a co-creative poetry system that makes people feel that they have “creative superpowers”? Presented at HAI-GEN Workshop.
- Oliveira, H. G. (2012). PoeTryMe: a versatile platform for poetry generation. *Computational Creativity, Concept Invention, and General Intelligence*, 1, p. 21.

- Oliveira, H. G. (2017). A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, Association for Computational Linguistics, pp. 11–20.
- Oliveira, H. G. (2021). Exploring a masked language model for creative text transformation. In *Proceedings of the Twelfth International Conference on Computational Creativity (ICCC)*, Mexico City, Mexico (Virtual): Association for Computational Creativity (ACC), pp. 62–71.
- Oliveira, H. G., Mendes, T. and Boavida, A. (2017). Co-poetryme: a co-creative interface for the composition of poetry. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, Association for Computational Linguistics, pp. 70–71.
- Oliveira, H. G., Hervás, R., Díaz, A. and Gervás, P. (2017). Multilingual extension and evaluation of a poetry generator. *Natural Language Engineering*, 23(6), pp. 929–967.
- OpenAI (2022). Introducing ChatGPT. <https://openai.com/blog/chatgpt/>, accessed: 2024-12-17.
- OpenAI (2023a). OpenAI API documentation. <https://platform.openai.com/docs/introduction>, accessed: 2023-01-26.
- OpenAI (2023b). OpenAI pricing. <https://openai.com/api/pricing/>, accessed: 2023-01-20.
- OpenAI (2024). GPT-4o. <https://openai.com/index/hello-gpt-4o/>, accessed: 2024-05-13.
- Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 27730–27744.
- Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.
- Pardinas, R., Huang, G., Vazquez, D. and Piché, A. (2023). Leveraging human preferences to master poetry. In *Proceedings of The AAAI-23 Workshop on Creative AI Across Modalities (WCAIM)*, AAAI Press.

- Pennington, J., Socher, R. and Manning, C. D. (2014). GloVe: Global Vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.
- Pérez y Pérez, R. and Sharples, M. (2004). Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-Based Systems*, 17(1), pp. 15–29.
- Peters, M. E. et al. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1 (Long Papers), New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.
- Pichai, S. and Hassabis, D. (2025). Introducing Gemini: Our largest and most capable AI model. <https://blog.google/technology/ai/google-gemini-ai/#introducing-gemini>, accessed: 2025-01-05.
- Plank, B., Søgaard, A. and Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 2 (Short Papers), Berlin, Germany: Association for Computational Linguistics, pp. 412–418.
- Poetry Foundation (2023). About Poetry Magazine. <https://www.poetryfoundation.org/poetrymagazine/about>, accessed: 2023-12-01.
- Poetry Foundation (2024). Poetry Foundation. <https://www.poetryfoundation.org/>, accessed: 2024-05-11.
- Poetry Free-For-All (2024). Newbie Stretching Room. <https://www.everypoet.org/pffa/>, accessed: 2024-04-21.
- Poetry.com (2023). Poems by genre. <https://www.poetry.com/poems-by-genre>, accessed: 2023-05-07.
- Porter, B. and Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1), p. 26133.
- Prince, S. J. (2023). *Understanding Deep Learning*. Cambridge, MA: MIT Press.
- Project Gutenberg (2024). Project Gutenberg. <http://gutenberg.org/>, accessed: 2024-06-08.

- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018). Improving language understanding by generative pre-training. OpenAI Blog, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- Radford, A. et al. (2019). Language models are unsupervised multitask learners. OpenAI Blog, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- Rahman, F. and Manurung, R. (2011). Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Mexico City, Mexico: Association for Computational Creativity (ACC), pp. 4–9.
- Rahmeh, H. (2023). Digital verses versus inked poetry: Exploring readers' response to AI-generated and human-authored sonnets. *Scholarly International Journal of Linguistics and Literature*, 6(9), pp. 372–382.
- Rashel, F. and Manurung, R. (2014). Pemuisi: a constraint satisfaction-based generator of topical Indonesian poetry. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Ljubljana, Slovenia: Association for Computational Creativity (ACC), pp. 82–90.
- Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, 44(3), pp. 393–401.
- Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan*, 42(7), pp. 305–310.
- Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17, pp. 67–99.
- Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning internal representations by error propagation. *Nature*, 323, pp. 533–536.
- Sawicki, P. (2023). Fine-tuning GPT-3 for Poetry Generation and Evaluation: Dataset and Code. <https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation>.
- Sawicki, P., Grzes, M., Jordanous, A., Brown, D. and Peeperkorn, M. (2022). Training GPT-2 to represent two Romantic-era authors: challenges, evaluations and pitfalls. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Bozen-Bolzano, Italy: Association for Computational Creativity (ACC), pp. 35–43.

- Sawicki, P. et al. (2023a). Bits of Grass: Does GPT already know how to write like Whitman? In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Waterloo, Ontario, Canada: Association for Computational Creativity (ACC), pp. 317–321.
- Sawicki, P. et al. (2023b). On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Waterloo, Ontario, Canada: Association for Computational Creativity (ACC), pp. 10–19.
- Sawicki, P., Grzes, M., Brown, D. and Goes, F. (2025). Can large language models outperform non-experts in poetry evaluation? a comparative study using the consensual assessment technique. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China: Association for Computational Linguistics, pp. 31889–31906.
- Scao, T. L. et al. (2023). BLOOM: A 176B-parameter open-access multilingual language model. *HAL preprint*, available at: <https://inria.hal.science/hal-03850124>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint*, 1707.06347.
- Shao, Y., Shao, T., Wang, M., Wang, P. and Gao, J. (2021). A sentiment and style controllable approach for Chinese poetry generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, ACM, pp. 4784–4788.
- Sheng, E. and Uthus, D. (2020). Investigating societal biases in a poetry composition system. *arXiv preprint*, 2011.02686.
- Shepperd, N. (2021). nshepperd/gpt-2: Code for the paper “Language models are unsupervised multitask learners”. <https://github.com/nshepperd/gpt-2>, accessed: 2022-01-30.
- Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), pp. 420–438.
- Sutskever, I., Vinyals, O. and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3104–3112.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press, 2nd edn.
- Taori, R. et al. (2023). Alpaca: A strong, replicable instruction-following model. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, accessed: 2024-12-29.

- The Vicuna Team (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. <https://lmsys.org/blog/2023-03-30-vicuna/>, accessed: 2024-12-29.
- Tobing, B. C. L. and Manurung, R. (2015). A chart generation system for topical metrical poetry. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Park City, Utah, USA: Association for Computational Creativity (ACC), pp. 308–314.
- Toivanen, J. M., Järvisalo, M., Toivonen, H. et al. (2013). Harnessing constraint programming for poetry composition. In *Proceedings of The Fourth International Conference on Computational Creativity (ICCC)*, Sydney, Australia: The University of Sydney, pp. 160–167.
- Touvron, H. et al. (2023a). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, 2307.09288.
- Touvron, H. et al. (2023b). Llama: Open and efficient foundation language models. *arXiv preprint*, 2302.13971.
- Uthus, D., Voitovich, M. and Mical, R. (2022). Augmenting poetry composition with Verse by Verse. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, Hybrid: Seattle, Washington + Online*: Association for Computational Linguistics, pp. 18–26.
- Uthus, D., Voitovich, M., Mical, R. and Kurzweil, R. (2019). First steps towards collaborative poetry generation. In *Proceedings of the NeurIPS Workshop on Machine Learning for Creativity and Design (ML4CD)*, vol. 3, a NeurIPS 2019 Workshop.
- Vaswani, A. et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008.
- Veale, T. (2013). Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit. In *Proceedings of the International Conference on Computational Creativity (ICCC)*, Sydney, Australia: Association for Computational Creativity (ACC), pp. 152–159.
- Ventura, D. (2016). Mere generation: Essential barometer or dated concept. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC)*, Paris, France: Sony CSL, pp. 17–24.
- Vuong, O. (2013). DetoNation. <https://www.poetryfoundation.org/poetrymagazine/poems/56768/detonation>, published in Poetry Magazine. Accessed: 2024-09-13.

- Wang, B. and Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model. GitHub repository. <https://github.com/kingoflolz/mesh-transformer-jax>.
- Wang, Z. et al. (2016). Chinese poetry generation with planning based neural network. *arXiv preprint*, 1610.09889.
- Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 24824–24837.
- Wentink, M. (2023). *Creating and evaluating a lyrics generator specialized in rap lyrics with a high rhyme density*. BSc thesis, University of Twente.
- Wesek, A. (2021). *A comprehensive study of state-of-the-art word embedding algorithms for natural language*. Master's thesis, University of Kent (UK).
- Whitman, W. (1865). Spirit whose work is done. https://whitmanarchive.org/item/ppp.00270_00412, from Walt Whitman Archive. ID: ppp.00270.00412, Washington City, 1865. Accessed: 2023-11-23.
- Whitman, W. (n.d.). Who learns my lesson complete? https://whitmanarchive.org/item/ppp.00707_00925, from Walt Whitman Archive. ID: ppp.00707.00925. Accessed: 2023-11-23.
- Wiggins, G. A. (2006). A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems*, 19(7), pp. 449–458.
- Williams, P. B. (2013). Of darker ceremonies. <https://www.poetryfoundation.org/poetrymagazine/poems/56573/of-darker-ceremonies>, published in Poetry Magazine. Accessed: 2024-09-13.
- Wöckener, J. et al. (2021). End-to-end style-conditioned poetry generation: what does it take to learn from examples alone? In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL)*, Association for Computational Linguistics, pp. 57–66.
- Wolf, T. et al. (2019). HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint*, 1910.03771.
- Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2145–2158.

- Yang, D., Zhou, Y., Zhang, Z., Li, T. J.-J. and LC, R. (2022). AI as an active writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the ACM IUI Workshops (IUI Workshops)*, vol. 10, CEUR-WS.org, pp. 1–11.
- Yang, Z. et al. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, Curran Associates, Inc.
- Yu, C., Zang, L., Wang, J., Zhuang, C. and Gu, J. (2024). Charpoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 3: System Demonstrations)*, Association for Computational Linguistics, pp. 315–325.
- Yu, L., Zhang, W., Wang, J. and Yu, Y. (2017). SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 31, AAAI Press.
- Zhang, R. and Eger, S. (2024). LLM-based multi-agent poetry generation in non-cooperative environments. *arXiv preprint*, 2409.03659.
- Zhang, S. et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint*, 2205.01068.
- Zhang, X. and Lapata, M. (2014). Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, pp. 670–680.
- Zhipeng, G. et al. (2019). Jiuge: A human-machine collaborative Chinese classical poetry generation system. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, Association for Computational Linguistics, pp. 25–30.
- Zugarini, A., Melacci, S. and Maggini, M. (2019). Neural poetry: Learning to generate poems using syllables. In *Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series (ICANN)*, Springer, pp. 313–325.