

DISSERTATION

ON THE CREATIVE APPLICATION OF LARGE LANGUAGE MODELS

Max Peeperkorn

Submitted in fulfilment of the degree of

DOCTOR OF PHILOSOPHY

in the subject of

COMPUTER SCIENCE

SCHOOL OF COMPUTING

UNIVERSITY OF KENT

OCTOBER 2026

Supervisors

Dr. Anna Jordanous

Prof. Daniel G. Brown

(University of Waterloo, Canada)

Supervisory Chair

Dr. Marek Grześ

Examiners

Dr. Maria Teresa Llano

(University of Sussex, United Kingdom)

Dr. Kazjon Grace

(University of Sydney, Australia)

Dr. Ioanna Giorgi

Copyright © 2025 Max Peeperkorn

This research in this dissertation is supported by a Google PhD Fellowship in Natural Language Processing, het Cultuurfonds, het Hendrik Mullerfonds, het Vreedefonds, and the Sir Richard Stapley Educational Trust.

Dedicated to Mia

Acknowledgements

The road to a PhD in Computer Science has not been obvious for me at all. After a degree in Graphic Design, I pursued a Master's degree in the Media Technology programme at Leiden University. Through my Master's thesis supervisor, Rob Saunders, I learned about a possible PhD at the University of Kent, and I am thankful to Rob for connecting me with Anna Jordanous.

Anna, from our first meeting, I immediately knew I was in the right place. I feel fortunate to have had you as my supervisor. You were always available to share your wisdom, discuss tough research decisions, and support me through the inevitable stressful moments. Pursuing a PhD at a time when Large Language Models are challenging the limits of what we thought a machine could achieve has been intense and demanding. Still, I am incredibly thankful for the freedom you gave me to explore my own research interests, even when they distracted me from the overall goal. It has been a wonderful journey; thank you!

Dan Brown, thank you for all the insightful perspectives, notes, and conversations in these evolving times; I have learned a great deal from you. I am happy you were able to join the supervisory team in a formal role; you have been an invaluable part of this journey, for which I am grateful. I remember fondly when we all got to spend time together in person at ICCV'23 in Waterloo and our visit to Romney Marsh.

In many ways, a PhD is a lonely endeavour. I consider myself incredibly fortunate that my dear friend, Tom Kouwenhoven, whom I met during my Master's degree, was also pursuing a PhD. Going through this experience together has made it more enjoyable in every imaginable way. Tom, you wondered whether you helped me as much as you believed I helped you, and of course you did; having someone to talk to about research, no matter the hour, has meant more than I can say. We could discuss anything, even the smallest detail or the most mundane topic. I hope we can keep our collaboration and friendship going, even if we ultimately end up in different research areas or jobs.

I want to thank my external examiners, Maria Teresa Llano and Kazjon Grace, and my internal examiner, Ioanna Giorgi, for their careful reading and insightful, detailed feedback. I would also like to thank Dominique Chu and Marek Grześ, who, as supervisory chairs at different stages of my PhD, were always open to giving advice and guidance.

On my first day, I was assigned a desk that was too small to work at, and somehow I ended up in a room with the students and postdocs researching Programming Languages. That same day, I remember being asked, “Do you drink beer?” and being promptly invited to the pub. Marco, I would like to thank you for your friendship over the past years at work and outside.

Ending up in a Programming Languages office turned out to be a blessing. I have learned about a great many things that might otherwise never have crossed my path, such as monads and monoids. Learning about different research areas through my peers in Artificial Intelligence, Programming Languages, and Cybersecurity has been a privilege. For that, I want to thank those with whom I shared this journey over the past years: Hrutvik, Tori, Jay, Rachel, Paulo, Piotr, Lisa, Lee, Aidin, Joshua, Andrew, Mike, and all my other fellow PhD students who have already moved on, and those who have started since.

I am thankful for all my friends and family who have kept in touch since we moved abroad. In particular, I thank my parents, who have always supported me in finding my own way in life, and my brother and sister, who are always there whenever I need them.

Finally, Alice, I am so very lucky that you wanted to take on this adventure abroad with me. There is simply no way to thank you for your endless support and belief in me. You are the keystone of my life.

On the Creative Application of Large Language Models

Max Peeperkorn 

School of Computing

University of Kent, Canterbury, United Kingdom

mp770@kent.ac.uk

Abstract

Large language models (LLMs) have become a common sight in day-to-day activities and have found use for a wide range of tasks, many of them of a creative nature. While not perfect, these models have enabled automation of tasks that were previously difficult to automate, such as generating fluent narratives or using them as creative evaluators. But to what extent can pure large language models perform creative tasks? This dissertation explores and evaluates the use of LLMs without augmentations for creative applications, specifically in the domain of narrative generation.

The first contribution of this dissertation is an analysis of LLMs and their creative potential, including evaluation using three established CC evaluation frameworks: “Mere Generation”, The Creative Tripod, and FACE. I argue that it is inappropriate to evaluate LLMs using these frameworks, primarily because their “process” is simple, yet they can perform complex tasks. Moreover, these frameworks often require explicitly modelled features to argue for the creativity of the system. However, for LLMs, their abilities are implicit in their learned parameters.

Secondly, I shift focus to empirical evaluation to investigate a simple yet common claim in popular media and academic literature: is temperature the universal creativity parameter of pure large language models? I have designed a novel experimental method based on theories from cognitive science, and reframed the temperature zero story as the exemplar story for a given prompt. This exemplar enables comparison in a human evaluation experiment to understand the impact of the temperature hyperparameter. The results indicate that this claim is far more nuanced and weaker than is often presented. It appears

higher temperature enable more exploration by increasing the chance of generating more variety, but it does not allow access to a significantly larger slice of the probability distribution.

This observation raises questions regarding the output diversity of LLMs, and particularly, those trained for instruction following. The third contribution of this dissertation is two-fold. First, I investigate the gap in output diversity between instruction-tuned models and their corresponding base models. While this phenomenon has been observed before, I confirm its existence in the domain of narrative generation and demonstrate further findings on how intermediate fine-tuning steps impact the output diversity. Secondly, I present a novel decoding strategy, *conformative decoding*, motivated by the diversity gap that aims to reintroduce some of the lost output diversity. Conformative decoding works by mixing the next-token distribution of the base model into the next-token distribution of the instruction-tuned model. The computational results show a slight increase in output diversity. In contrast, the human evaluation experiment reveals no preference for either the baseline or conformative decoding. Assessing the diversity of a set of texts is a complex task, and the difference in diversity may be too subtle to detect or not meaningful to humans.

LLMs are powerful tools, and we have only begun understanding their creative potential. The findings in this dissertation show the limits of pure LLMs performing creative tasks and our ability to evaluate them using current methods. However, based on the findings, I believe there is much potential for future work on using multi-agent setups that view creativity as a social phenomenon.

Contents

List of Publications	x
1 Introduction	1
1.1 Topic of the Dissertation	1
1.2 Dissertation Design	6
1.2.1 Research Questions	6
1.2.2 Methods	11
1.2.3 Structure	13
2 Background	16
2.1 What is Creativity?	16
2.2 Large Language Models	19
2.2.1 Transformers	20
2.2.2 Decoding Strategies	25
2.2.3 Instruct Models and Preference Optimisation	30
2.3 Narrative Generation: Traditional Approaches and LLMs	32
2.3.1 Traditional Narrative Generation Systems	32
2.3.2 Narrative Generation using Large Language Models	33
2.4 Towards an Understanding of LLM Creativity	34
3 LLMs and CC Evaluation Frameworks	35
3.1 Introduction	35
3.2 Background	37
3.2.1 Transformer-based Language Models	37
3.2.2 LLMs and Creative Tasks	39

3.2.3	Creativity Evaluation	40
3.3	Computational Creativity Evaluation Frameworks	41
3.3.1	Mere Generation	41
3.3.2	The Creative Tripod	44
3.3.3	FACE/IDEA	46
3.4	Creativity Analysis	48
3.4.1	Scope: The Creative System	49
3.4.2	Methodology	49
3.4.3	Analysis	50
3.5	Discussion	58
3.6	Conclusion	60
4	Is Temperature the Universal Creativity Parameter?	62
4.1	Introduction	62
4.2	Background	64
4.2.1	Conditions for Creativity in Narrative Generation	64
4.2.2	The Temperature Parameter	66
4.2.3	Probing Large Language Models	67
4.2.4	Prototypes and Exemplars	67
4.3	The Exemplar Story	71
4.3.1	Identifying the Exemplar	71
4.3.2	Generating the Exemplar Story	72
4.4	Case Study: “Write a story.”	73
4.4.1	Computational Analysis	73
4.4.2	Experimental Analysis with Human Evaluation	78
4.5	Discussion	87
4.6	Conclusion	88
5	Conformative Decoding: Minding the Diversity Gap	90
5.1	Introduction	90
5.2	Background	92
5.2.1	What is diversity?	92

5.2.2	Measuring Quality and Diversity	93
5.2.3	Improving the Output Diversity of LLMs	99
5.3	The Diversity Gap	100
5.3.1	Experimental Setup	100
5.3.2	Results	105
5.4	Conformative Decoding	109
5.4.1	Decoding Strategy	109
5.4.2	Experiment Setup	110
5.4.3	Results	111
5.4.4	Human Evaluation	111
5.5	Discussion and Limitations	115
5.5.1	Why does Llama 3.1 8B behave differently?	117
5.5.2	Challenges for Human Evaluations of Diversity	118
5.6	Conclusion	120
6	Conclusions	121
6.1	Answers to Research Questions	122
6.2	Reflection	125
6.2.1	Limitations	125
6.2.2	Discussion	128
6.3	Have Large Language Models changed?	132
6.3.1	Future Work	133
	Bibliography	137
A	Supplementary Materials for Chapter 3	189
B	Supplementary Materials for Chapter 4	192
C	Supplementary Materials for Chapter 5	197

List of Publications

This dissertation comprises the work presented in the original papers listed below.

1. Peeperkorn, M., Brown, D. G., & Jordanous, A. (2023). On characterizations of large language models and creativity evaluation. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 143–147). Association for Computational Creativity
2. Peeperkorn, M., Kouwenhoven, T., Brown, D. G., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models? In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 226–235). Association for Computational Creativity
3. Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2025). Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models [Preprint, Under review]. <https://arxiv.org/abs/2507.20956>

~

During the PhD, I have worked on related research, and while these papers do not represent direct contributions, they have shaped and informed the work presented in this dissertation.

1. Peeperkorn, M. (2022). Artificial Creative Societies: Adaption, Intention, and Evaluation. *Creativity and Cognition*, 704–707. <https://doi.org/10.1145/3527927.3533728>
2. Sawicki, P., Grześ, M., Jordanous, A., Brown, D. G., & Peeperkorn, M. (2022). Training GPT-2 to represent two romantic-era authors: Challenges, evaluations and pitfalls. In M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz & T. Veale (Eds.), *Proceedings of the 13th international conference on computational creativity* (pp. 34–43)
3. Brown, D. G., & Peeperkorn, M. (2023). Reviewing, creativity, and algorithmic information theory. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 133–147)
4. Sawicki, P., Grześ, M., Goes, F., Brown, D. G., Peeperkorn, M., & Khatun, A. (2023a). Bits of grass: Does GPT already know how to write like whitman? In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 317–321)
5. Sawicki, P., Grześ, M., Goes, F., Brown, D. G., Peeperkorn, M., Khatun, A., & Paraskevopoulou, S. (2023b). On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 10–19)

6. Kouwenhoven, T., Peeperkorn, M., Van Dijk, B., & Verhoef, T. (2024). The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke & Y. Oseki (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 57–71). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.cmcl-1.5>
7. Kouwenhoven, T., Peeperkorn, M., & Verhoef, T. (2025). Searching for structure: Investigating emergent communication with large language models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 9977–9991). Association for Computational Linguistics. <https://aclanthology.org/2025.coling-main.667/>
8. Kouwenhoven, T., Peeperkorn, M., de Kleijn, R., & Verhoef, T. (2025). Shaping shared languages: Human and large language models' inductive biases in emergent communication. In J. Kwok (Ed.), *Proceedings of the thirty-fourth international joint conference on artificial intelligence, IJCAI-25* (pp. 10298–10306). International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2025/1144>



Chapter 1

Introduction

This dissertation aims to further our understanding of large language models applied to creative purposes using theory, methods, and approaches from the fields of computational creativity, generative artificial intelligence, computational linguistics, philosophy, and cognitive psychology. I present a study of the creative abilities and utility of large language models. The dissertation comprises an overview of what large language models are and a creativity evaluation using three computational creativity frameworks, followed by empirical studies and a novel method to generate more diverse outputs. The focus is on evaluating the creativity of large language models by their generated artefacts along the axes of novelty, typicality, quality, and diversity.

1.1 Topic of the Dissertation

Computer science is about investigating and determining what tasks can be automated; the goal is not just to build a machine, it is to build and then study it (Brooks, 1996). In this dissertation, I study such a machine, the large language model, to determine if it is useful for automating creative tasks, in particular, narrative generation tasks.

Creativity is a profoundly complex phenomenon that has a variety of definitions depending on the context of the creative artefact and one's perspective. Although the exact relationship between creativity and human intelligence is debated, there is clearly a connection (Sternberg & O'Hara, 1998; Kaufman & Plucker, 2011). The creative mind has given humanity remarkable tools and has led to many breakthroughs, both scientific and cultural. Creativity is deeply rooted in social and cultural processes (Amabile, 1982; Csikszentmihalyi, 1988; Vygotsky, 2004), making it even more elusive and particularly hard to capture in computational processes. Mechanising creativity and making machines that make art and music is not new; it has been on the minds of people for centuries. For example, around 1800, a Swiss mechanic, Henri Maillardet, built an automaton that could draw pictures and write short verses produced from its associative memory.¹ To this day, artificial creativity is an important challenge and interest in AI research (e.g., Novikov et al., 2025).

But what is a creative task? Creativity research in Psychology distinguishes between everyday creativity (little-c) and eminent creativity (big-C; Kaufman & Beghetto, 2009).² Little-c is about everyday problem-solving: the small tasks that we all encounter on a daily basis, such as improvising a meal from whatever is left in the fridge, or finding a workaround for a broken tool. Big-C creative tasks, by contrast, are about invention and innovation, like scientific research, developing new technologies, art, poetry, music, and so on (Kaufman & Beghetto, 2009). While little-c and Big-C tasks can be challenging for machines, it is the latter that most AI researchers are interested in.

The breakthrough of Large Language Models (LLMs) and the transformer architecture (Vaswani et al., 2017), and then the release of ChatGPT (GPT-3.5) in 2022, led to a substantial leap in what AI models could do. The fluency and quality of their outputs are unparalleled compared to previous techniques and methods (Radford et al., 2018). In particular, the ability to *instruct* the LLM is powerful and productive.

¹See <https://web.archive.org/web/20070821154911/http://www.fi.edu/pieces/knox/automaton/index.html>

²This is part of the four-C Model of Creativity by Kaufman and Beghetto (2009) that also introduces mini-c and Pro-c for the personal/developmental aspect of creativity and professional creativity, respectively.

LLMs have found use in many domains traditionally associated with creativity, and the question that naturally arises is whether these remarkable LLMs and their products are, in fact, creative. Given this recent surge in generative AI, are we getting closer to human-level artificial creativity?

This dissertation is centred in the interdisciplinary field of Computational Creativity (CC), a subfield of artificial intelligence that intersects with other research areas, such as computational linguistics, cognitive psychology, philosophy, and the arts. CC research, according to the Association for Computational Creativity (ACC), is concerned with three challenges (definition by ACC, quoted in Jordanous, 2014):

1. to construct a program or computer capable of human-level creativity;
2. to better understand human creativity and to formulate an algorithmic perspective on creative behaviour in humans;
3. to design programs that can enhance human creativity without necessarily being creative themselves.

The first challenge is one of engineering, the second is scientific in nature, using the machine as an instrument to say something about creativity itself, and the third is one of design, where CC borders on human-computer interaction.

These challenges are formulated a little awkwardly, as they suggest that CC is purely focused on generation, and on designing processes that mimic human creativity. However, a more recent view, which sits outside those challenges, is that machine creativity and human creativity are different, because their nature is different (Colton et al., 2020). In this frame, CC is also about exploring non-anthropomorphic perspectives of creativity (Guckelsberger, Salge & Colton, 2017), about defining formal mappings of creative processes (Lahikainen, Ady & Guckelsberger, 2024), and about algorithmic perspectives of creativity (Mondol & Brown, 2021a; Mondol & Brown, 2021b; Brown & Peeperkorn, 2023). Furthermore, over the past decades, it has become clear that evaluation of the generation process as well as the evaluation of artefacts, if not more (e.g., Colton, 2008;

Colton, Charnley & Pease, 2011; Jordanous, 2012). Although, as we shall see in Chapter 3 this non-trivial for LLMs.

Colton and Wiggins (2012) provide a foundational definition of CC: *“The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative.”* The notion of responsibilities is to separate the CC system from creative productivity tools. The domain of Creativity Support Tools (CSTs; Shneiderman, 2002) is often contrasted with CC because such tools are not expected to exhibit any creative agency, and as a consequence, have no responsibility in the creative process. CSTs have their own evaluation traditions, most prominently the Creativity Support Index (CSI; Cherry & Latulipe, 2014), which measures how well a tool supports the creativity of its user. This dissertation uses CC evaluation frameworks rather than instruments of this kind, because the question here is whether the system itself is creative, while the CSI evaluates the interaction between a user and a tool, and locates the creative agency with the user.

The creativity of LLMs is by means that assessing LLM creativity by means of tests from psychology, such as the divergent association task (DAT; Olson et al., 2021), the alternative uses task (AUT; Guilford, 1967) and the Torrance Tests of Creative Thinking (TTCT; Torrance, 1966). But that is beyond the scope of this chapter, as these do not focus on creative artefacts or processes, but rather uses specific human tests as a proxy to establish potential creative capabilities.

Responsibility, however, is not something a system simply has or lacks, as it can originate from a broad range of sources, for example, background knowledge, social context, motivation, and materials (Johnson, 2014). Instead, Johnson (2014) argues that a CC system should play a *role* in producing behaviours that an unbiased observer would deem creative. This opens up possibilities for the system to take on different roles: it may generate, it may evaluate, it may collaborate, or it may do several of these at once. The term “unbiased observers” also emphasises evaluation as a key component of CC research. This aligns with the broad categorisation of CC research along two axes: generation and evaluation (McGraw & Hofstadter, 1993; Pérez y Pérez & Sharples, 2001;

Gervás, Pérez & Sosa, 2007).

The emphasis in this dissertation is on the evaluation of creative behaviours in LLMs, leaving generation with a modest role (Chapter 5 features modifications to the generation process). The probabilistic nature of LLMs makes evaluation challenging for the kind of questions this dissertation aims to address. The main difficulty is reproducibility, as even a newline character in a prompt can dramatically influence the generated output (Hedderich et al., 2025). This means that in this dissertation, I leave the “creative process” up to the LLM: keeping the prompt as simple as possible, and setting hyperparameters that are not the focus of investigation to sensible standard values. As such, the focus is strictly not on prompt engineering. Moreover, the LLMs examined in this dissertation are what I consider pure LLMs, defined as follows:

Definition 1 (Pure Large Language Model). *An autoregressive transformer-based Large Language Model and a decoding strategy, without any further augmentation.*

A large language model is autoregressive because it predicts the next token based on the previous tokens in the sequence, meaning it uses the prefix (the input and the output so far) to generate subsequent tokens until an end-sequence token is generated. The distinction between a decoding strategy and an augmentation is one of *where* the intervention happens. A decoding strategy acts on the next-token distribution: it reshapes what the model is about to sample to improve the model’s behaviour in some principled way. An augmentation acts around the model, introducing procedures that run outside that pass and that take on part of the work themselves with the goal of guiding the model to more factually accurate outputs, or to converge on a desired result. Examples of augmentations are: Retrieval Augmented Generation (Lewis et al., 2020) supplies knowledge the model does not hold, a multi-agent simulation (Park et al., 2023) supplies a social process, an evolutionary-algorithm-like setup (Romera-Paredes et al., 2024; Novikov et al., 2025) supplies selection and iteration, and mixture-of-experts via a router network to select the best fit for the current query (Shazeer et al., 2017; Jiang et al., 2024). Truncation (Welleck et al., 2024) is the minimal decoding strategy, and a necessary one: the softmax function over-approximates the probabilities of the long tail of unlikely

tokens (the softmax bottleneck; Yang et al., 2018), so sampling from the untruncated distribution yields degenerate text. Conformative Decoding, presented in Chapter 5, is a more elaborate case, but it remains within this definition for the same reason: like contrastive decoding (Li et al., 2023; O’Brien & Lewis, 2023), it consults a second model to reweight the distribution, and no procedure is added around the generation process. In the next section, I present the research questions addressed by this dissertation and its structure.

1.2 Dissertation Design

This section presents the research questions, the contributions, and the structure of the dissertation.

1.2.1 Research Questions

LLMs require significant (computational) resources to train (Zhuang et al., 2023). Therefore, I opted to focus on existing open-source and open-weight models and study them as-is to determine their creative potential and limitations. Thus, the main research question is as follows:

Main Research Question

To what extent can pure large language models perform creative tasks?

This is a broad question, and to pursue an answer, the main question breaks down into four research questions that each address an important aspect of assessing and understanding the creative abilities of a pure large language model.

Research Question 1 (Chapter 3)

Can we evaluate the creativity of pure large language models using existing CC frameworks?

LLMs are challenging what we believed could be automated, and creative tasks are no exception. Therefore, it is essential to start by positioning where we are using existing frameworks establishing creativity. I employ three CC evaluation frameworks: “Mere Generation” (Ventura, 2016), the Creative Tripod (Colton, 2008) and FACE (Colton, Charnley & Pease, 2011; Pease & Colton, 2011a) to address the question if LLMs can be considered creative. While this is not an exhaustive list of evaluation frameworks, they are chosen as they each address a relevant and different perspective, namely: knowledge representation, external perception and creative acts, respectively. Besides understanding if LLMs are more than “mere” generators, if they can be perceived to be creative, and what kind of creative acts they can perform, the aim is to understand the applicability of these frameworks, and to what extent we must rely on empirical evaluation methods and human evaluation studies.

Contribution While the analysis shows that for all three frameworks a pure LLM (Definition 1) can generally be seen as creative under these systems, they all fail in some way. Each framework relies on pointing out a (family of) function(s) that are part of the system to argue for creative behaviour, which is hard to probe reliably for LLMs as these functions, if they are performed, are performed implicitly. Furthermore, LLMs produce output of high quality and so it becomes difficult to detect whether it is really the result of their creativity. The consequence is methodological, and it shapes the rest of the dissertation: what is responsible for the creativity of LLMs cannot be argued from the system, so it must be observed empirically.

Research Question 2 (Chapter 4)

Is temperature the universal creativity parameter of pure large language models?

Research Question 1 showed that CC evaluation frameworks are insufficient for fully assessing what is responsible for the creative behaviours of LLMs, and subsequently guided us towards an empirical methodology for the remainder of this dissertation. In Chapter 4, I address a key question based on a commonly held assumption that the creative abilities of LLMs can be controlled by means of the temperature hyperparameter (Manjavacas et al., 2017; Roemmele & Gordon, 2018; Chen & Ding, 2023). The temperature parameter controls the amount of uncertainty or randomness in the generation process of LLMs. This refers back to a long-standing debate in the CC community about whether a random process is creative or not. There have been arguments against (Veale, 2015; Ventura, 2016) and in favour (Simonton, 1999) of this idea. Temperature is a simple parameter whose value influences the output variation of LLMs (see Chapter 4, Section 4.2.2). Hence, a better understanding of its effect on the creative behaviours of LLMs is an essential step towards the central research question.

Contribution I address the question with a simple, controlled experiment using narrative generation, with a single model and a single prompt so that temperature is the only variable that moves. Within this setup, higher temperatures do lead to more exploratory behaviour in the embedding space, but only for a limited number of samples, and they do not enable access to a large slice of the overall probability distribution. The human evaluation shows a weak positive correlation between temperature and novelty, and a moderately negative correlation with coherence, which is a trade-off rather than a control. The claim under test is a universal one, that temperature is *the* creativity parameter of LLMs, and a single controlled setup is enough to show that it does not hold universally, but it is not sufficient to establish the effect of temperature in general. The chapter also contributes a methodology, built on the exemplar object, for probing hyperparameter effects while limiting the variance introduced by the prompt.

Research Question 3 (Chapter 5)

Does instruction-tuning reduce the output diversity of pure large language models for a narrative generation task?

Since temperature influences the output variation a further question is how much variety can a model produce in the first place. Prior work has shown that instruction tuning generally reduces the output diversity of LLMs (Hämäläinen, Tavast & Kunnari, 2023; Chu, Chen & Nakayama, 2024; Kirk et al., 2024; Le Bronnec et al., 2024; Shypula et al., 2025). Since diversity is a key component of creativity (Jordanous & Keller, 2016), this loss of diversity is arguably an undesirable effect for creative applications of large language models. This question, addressed in Chapter 5, tests for the same phenomenon on a simple narrative generation task using a subset of the Writing Prompts dataset (Fan, Lewis & Dauphin, 2018). I investigate the diversity gap through a computational analysis using standard metrics for assessing generative AI models (Kynkäänniemi et al., 2019; Pillutla et al., 2021) and metrics designed with diversity in mind: the Vendi Score (Friedman & Dieng, 2023) and Truncated Entropy (Ibarrola, Lawton & Grace, 2024). Instruction tuning consists of multiple phases, and each may contribute differently to a lower output diversity. Hence, for models that made the intermediate checkpoints available, I aim to assess each checkpoint for a loss of diversity.

Contribution The results provide additional evidence for the existence of the gap in output diversity between the instruct model and its base. Moreover, additional experiments show that Direct Preference Optimisation (Rafailov et al., 2023) has a large impact on the reduction in output diversity.

Research Question 4 (Chapter 5)

Can a pure large language model produce more diverse output, and maintain quality, by informing the instruction-tuned model using its associated base model?

Here, we investigate, based on the result of Research Question 3, if we can mitigate the diversity gap by using a novel decoding method inspired by this phenomenon. This question concerns two pure LLMs: an instruction-tuned model, and the *base* model it was fine-tuned from. Fine-tuning changes the weights, not the generation process, so both remain pure within Definition 1. The instruction-following ability of the LLM after

fine-tuning for human interaction makes it much more useful and productive than a base model. These base models need to be prompted on a completion basis, meaning they must be provided with examples to enable few-shot in-context learning (Brown et al., 2020). They also rely on pre-existing patterns in the training data. In practice, however, this is difficult, as one needs to find the right pre-existing patterns and have a good set of examples available. For creative tasks, we ideally require an instruction-following but diversity-preserving generator to produce the desired artefacts. In Chapter 5, motivated by the findings on the previous research question, I present a decoding strategy, *Conformative Decoding*, that reintroduces information from the base model in the generation process with the goal of improving output diversity. I assess its effectiveness in the same experimental setup as for Research Question 3, followed by a human evaluation experiment to validate the computational analysis.

Contribution I present a novel decoding strategy that reintroduces output diversity by having the instruct model conform to its base. The computational results overall show clear increases in diversity using Conformative Decoding, indicating that from an algorithmic perspective the method is proving effective. On the other, the human evaluation for stories generated for OLMo 2 7B 1124, showed no consensus among human participants on either diversity or quality; the difference might be too subtle to detect, and we require clearly defined domain-specific criteria for diversity and quality, as opposed to an open-ended evaluation providing only pointers for each criteria.

Summary

The Main Research Question asks to what extent pure LLMs can perform creative tasks, with Research Questions 1–4 responding to different parts of this overarching question. RQ1 sets the starting point, using CC evaluation frameworks to assess the creativity of an LLM can be established from the system itself; the answer determines whether the remaining questions can be approached analytically or must be approached empirically. RQ2 tests the most commonly assumed answer to the main question, that creativity is controlled by the temperature parameter, and establishes how far that control actually

Method	RQ1	RQ2	RQ3	RQ4
Framework Analysis	•			
Computational Analysis		•	•	•
Human Evaluation		•		•
Statistical Modelling		•	•	•
<i>Chapter</i>	3	4	5	5

Table 1.1: Overview of main methods used for each research question.

reaches. RQ3 asks what is lost when a model is made useful, since instruction tuning reduces output diversity, a key component of creativity (Jordanous & Keller, 2016), and RQ4 asks how much of that loss can be recovered from within the decoding process. Together, they establish what to expect creatively from a pure LLM, and where its limits come from.

Statement

The summary above sets out how the questions connect, leading to the following statement: the dissertation concludes that a pure LLM is a capable generator, but not, on its own, a complete creative system. This is not an argument that LLMs are unrelated to creativity, but one about where their contribution sits: as a component of a larger system that carries part of the creative responsibility, in the sense of Colton and Wiggins (2012) and Johnson (2014), rather than as a creative system in its own right. This is the position I defend in Chapter 6.

1.2.2 Methods

The creativity of computational systems in CC is studied in various ways, combining analytical, empirical, and computational methods. The CC community has developed

frameworks to assess creative systems through various criteria from different perspectives, such as external perception (Colton, 2008) or how knowledge is represented (Ventura, 2016). CC evaluation frameworks are a means to assess creativity, but it is also common to evaluate creative systems by asking human participants to judge their artefacts on selected criteria (e.g., Ritchie, 2007). Since the focus of this dissertation is on LLMs, I use computational evaluation metrics common in assessing generative AI models (e.g., Kynkäänniemi et al., 2019; Pillutla et al., 2021; Friedman & Dieng, 2023; Ibarrola, Lawton & Grace, 2024). These methods usually rely on a ground truth that possesses the properties (for example, quality) under evaluation. Novel techniques that influence the behaviour of generative models are also commonly assessed via human evaluation, often in a direct comparison of two setups, with and without the novel technique (e.g., Meister et al., 2023; Finlayson et al., 2024).

In this dissertation, I adopt these methods to address the above research questions. Table 1.1 presents an overview of methods used for each research question and in each chapter. Below, I elaborate on each method, providing guidance to the reader.

- **Framework Analysis** refers to the application of CC evaluation frameworks to analyse the creative behaviours of large language models (RQ1).
- **Human Evaluation** refers to evaluation experiments involving participants to judge the creative output of the large language model on various criteria via questionnaires (RQ2, RQ4).
- **Computational Analysis** refers to using computational methods to assess various properties of large language models (RQ2, RQ3, RQ4).
- **Statistical Modelling** refers to the use of linear mixed effects models for hypothesis testing to address the random effects introduced by participants evaluating different subsets of creative artefacts (RQ2), hypothesis testing of computational methods (RQ3, RQ4), and a binomial test for the comparison task between two experimental setups (RQ4).

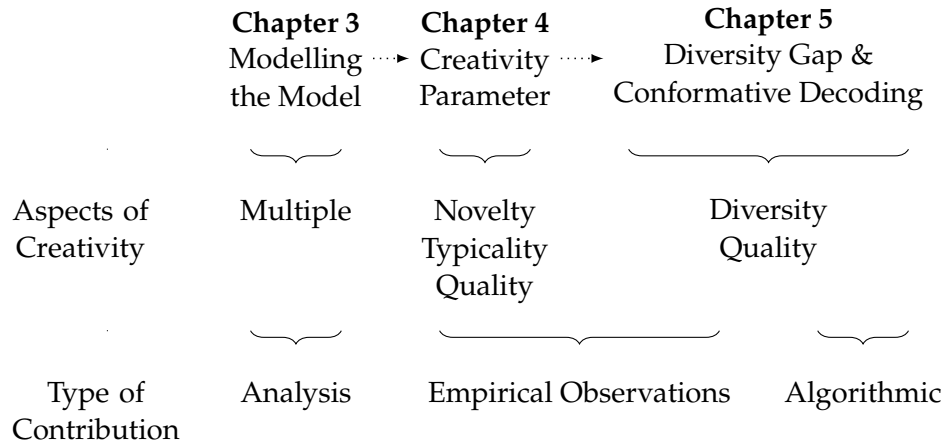


Figure 1.1: The structure of the dissertation constituting three content chapters, indicating their respective (creative) domains, what aspects of creativity are addressed, and the kind of contribution they provide.

1.2.3 Structure

Now that the research questions and methods are in place, this section is intended as a guide to the reader on how this dissertation is organised and its kind of contribution (the contributes are discussed above in Section 1.2.1). Figure 1.1 presents the scope and the kind of contribution of each chapter visually. The dissertation begins with a general background chapter that provides the core context for all chapters. However, the dissertation is designed so that each chapter can be read standalone, and each chapter includes a section that deepens its specific background. While this does introduce some redundancy, it provides a more focused presentation of work in this dissertation. Chapter 3 mainly uses CC evaluation frameworks, while Chapter 4 is influenced by theories from cognitive sciences, and Chapter 5 requires background on evaluation methods for diversity and quality used in generative AI.

Chapter 2 – Background

This chapter provides the *common background* that underpins all chapters. In particular, this chapter discusses LLMs and creativity, creativity evaluation, transformers and decoding strategies, and lastly, it gives a brief overview of the narrative generation domain that Chapters 4 and 5 have in common.

Chapter 3 – LLMs and CC Evaluation Frameworks

This *analysis* chapter assesses the creative abilities of pure LLMs (Definition 1) by means of three established CC evaluation frameworks by means of a systematic analysis. This chapter establishes a baseline understanding that informs the empirical research directions that are pursued in the remaining chapters.

Chapter 4 – Is Temperature the Universal Creativity Parameter?

This *experimental* chapter investigates a common understanding that the temperature hyperparameter of large language models controls their creativity. Drawing on theories from cognitive science, it presents a methodology to assess the effects of temperature applied to a narrative generation task.

Chapter 5 – Conformative Decoding: Minding the Diversity Gap

This *experimental* and *algorithmic* chapter focuses on diversity in outputs of LLMs and consists of two parts. First, it explores and verifies the existence of a difference in output diversity between the instruction-tuned LLM and its base model for a narrative generation task. Secondly, this chapter presents a decoding strategy, motivated by part one, to reintroduce some output diversity, which is evaluated using computational methods and a human evaluation based on direct comparison.

Chapter 6 – Conclusions

In this chapter, I present my current answers to the research questions. Moreover, I reflect on the research in this dissertation, providing discussions regarding the limitations, contributions, and directions for future work.

Appendices

The appendices contain examples of stories generated by the models and techniques presented in this dissertation, along with the questionnaires and other details regarding the human evaluation.

Chapter 2

Background

The aim of this chapter is to provide context and the prerequisites for the scope of the dissertation. I present a view on creativity, creativity evaluation, evaluation of generative systems, the transformer architecture, decoding strategies and a brief overview of computational narrative generation systems, and how they relate to Large Language Models (LLMs). Additionally, each individual chapter has a dedicated background with the essential relevant details.

2.1 What is Creativity?

Creativity is an essentially contested concept (Gallie, 1955): it undoubtedly exists, but we cannot agree on the exact conditions when something is creative. One of the first attempts to address this issue was put forward by Boden (1992) and has been pivotal for artificial creativity research. This framework was later formalised by (Wiggins, 2006a). Boden's framework first distinguishes between two creative behaviours, P- and H-creativity; psychological (or personal) and historical, respectively. A concept is H-creative if it has never been created before by any creator; on the other hand, it is only P-creative if the

concept already exists but is new to the creator. Secondly, Boden makes a distinction between two modes of creativity: exploratory and transformational. Key to these two notions is the idea of a metaphorical map of the mind, or a conceptual space (Gärdenfors, 2004). The conceptual space is what allows the discovery of new concepts by exploring uncharted territories, or if a concept is particularly novel and valuable, redrawing the map: transforming the conceptual space more fundamentally. There is an interesting parallel between conceptual spaces and the latent spaces learned by neural networks, which map objects into a high-dimensional vector space. These latent representations leverage the complex patterns in the training data, mapping semantically similar elements close together (DeMers & Cottrell, 1992; Kingma & Welling, 2014), making this a powerful technique to enable traversal and interpolation to potentially find novel objects (Peeperkorn et al., 2022). Boden also specifies a third mode of creativity, combinational creativity or conceptual blending, which simply entails combining concepts into new ones. Combining concepts is easy for machines, but very hard to do well; the challenge is to find combinations of high value (Boden, 1992).

So far, this only describes how we might find new concepts or transform spaces, and we have not defined creativity yet. The definition Boden uses is that something is creative when it is novel, surprising, and valuable. Surprise is more a property of the observer rather than the creator and derived from novelty (Wiggins, 2006a). Disregarding surprise, novelty and value is usually seen as the standard definition of creativity; this frame originates with Stein (1953) and Barron (1955). The two parts of this definition have many forms: novelty can be described as originality, unusualness, or uniqueness, and value as effectiveness, usefulness or quality (Runco & Jaeger, 2012).

The standard definition (novelty and value) is a high-level view towards assessing creativity. It emphasises artefact-based evaluation: judging creativity based on products, and implicitly judging its creator. However, creativity is a complex, multi-faceted phenomenon with numerous other perspectives (Walia, 2019), and many of these perspectives originate from psychology and are based on uniquely human traits. Nonetheless, psychological models of creativity, such as those that assess divergent thinking (Torrance, 1972; Dollinger, Urban & James, 2004), are increasingly relevant for Computational Creativity

(CC) research given that they can perform human-like tasks relatively fluently. Such psychological tests, like the divergent association task (DAT), alternative uses task (AUT) or the Torrance Tests of Creative Thinking (TTCT), are now being applied to evaluate the creative capabilities of AI models, like LLMs (e.g.; Stevenson et al., 2022; Góes et al., 2023b; Bellemare-Pepin et al., 2024). However, an open question is whether high performance on these psychological assessments indicates anything meaningful about an AI model’s ability to generate creative artefacts.

A useful frame for thinking about creative systems and their components is the four perspectives (4Ps) framework (Rhodes, 1961), which was expanded into CC by Jordanous (2016). The 4Ps break down into four aspects: the person (or producer), process, product, and press (the environment in which the creative artefact exists). A common view is that the creative behaviour of a creative system lies solely with the programmer (Turing, 1950), but this is not the kind of effect we are looking for. However, creative systems are designed precisely to leave room for unexpected behaviour and output (Colton, 2008). The machine of interest in this dissertation, the LLM, is no different, and can act beyond what the programmer intended due to its probabilistic nature (See section 2.2) and the intricate and complex patterns it learns from the vast amounts of training data. This is why it is interesting to assess whether the LLM can behave creatively.

The primary focus in this dissertation is on the creative product, using definitions and experimental setups appropriate for artefact-based evaluation (the product aspect of the 4Ps framework). The reason for my focus on artefact-based evaluation is that the *process* of pure LLMs is quite simple. Moreover, their black box nature makes it infeasible to assess this process and how it uses the complex interactions of the prompt and its parameters. While the prompt could be seen as a way to introduce more information about the *press*, this introduces much variability, complicating probing and generalisation from the empirical findings (as I will discuss in more detail in chapter 4).

There is work in explainable AI and the arts (e.g. Bryan-Kinns et al., 2021, 2024) that attempts to understand the creative process, but the LLMs examined in this dissertation are not special-purpose, meaning there are many more dimensions to take into account.

The transformer's attention mechanism can potentially be used to find patterns in prompts and the generation process, linking those to characteristics of the creative process, such as iteration, planning, and reflection. However, the creative process is a highly non-linear contextual endeavour, as exemplified in the 4Ps framework, which is somewhat at odds with the autoregressive LLM that is the subject of this dissertation. Hence, the focus on artefact-based evaluation.

In this dissertation, the LLM is evaluated in different ways using various criteria as necessary. Chapter 3 analyses LLMs in the broader perspective of creative aspects according to the three creativity evaluation frameworks: the Creative Tripod (Colton, 2008), FACE (Colton, Charnley & Pease, 2011; Pease & Colton, 2011a), and the Mere Generation frame (Ventura, 2016). In chapter 4, the creativity evaluation is done using the novelty, typicality, and quality criteria (these are discussed in more detail in Chapter 3 and Chapter 4). Following the findings of Chapter 4, I shift focus for Chapter 5 towards diversity, a key component of creativity (Jordanous & Keller, 2016), and keep an emphasis on quality.

2.2 Large Language Models

A language model is a model for natural language. Historically, there are three kinds of language models: formal models of language (Chomsky, 1965), statistical models (Manning & Schütze, 1999), and neural models (Jurafsky & Martin, 2025). Formal models of language purely use grammar (symbols and production rules) and are concerned with syntactic regularities. However, these models are less effective than statistical methods at showing an understanding of language (Church, 2011), which have turned out to be more successful for tasks such as text generation, machine translation, clustering, and classification (Manning & Schütze, 1999). However, simple statistical models like those based on n -grams struggled to generate fluent and coherent language due to the data sparsity problem. Words (or characters) combine in a huge number of ways, meaning that it is nearly impossible to get sufficient data to cover enough of the language space to

have a well-generalised statistical model.

Early neural approaches alleviated this problem by learning continuous vector representations, or embeddings (Mikolov et al., 2013). These embeddings capture semantic relationships, and are key to generating coherent output when processing sequential data. This led to the widespread adoption of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997). These networks maintain a hidden state that enables human-level sequence generation (Graves, 2013) and form the foundation of sequence-to-sequence modelling (Sutskever, Vinyals & Le, 2014). However, they can only model relatively short sequences, and their sequential nature is not easily parallelisable, which severely limits training capacity.

The introduction of the *transformer* (Vaswani et al., 2017), which allows parallel training, led to the next significant shift in the capabilities of language models (Radford et al., 2018). Transformer-based language models, usually consisting of billions of parameters, can easily produce fluent language, can accommodate very long sequences (e.g., OpenAI, 2023), and can learn rich semantic embedding spaces (Hewitt & Manning, 2019; Tenney, Das & Pavlick, 2019). Unsurprisingly, these models have become ubiquitous for natural language tasks, approaching or surpassing human-level performance (Kaplan et al., 2020) and can learn multiple tasks (Radford et al., 2019). In this dissertation, my focus is on autoregressive transformer-based language models for open-ended text generation. In the remainder of this section, I will explain the workings of the autoregressive transformer, decoding strategies, and instruction-tuning.

2.2.1 Transformers

The most successful LLMs for text generation have a decoder-only architecture. This kind of LLM is known as a Generative Pre-trained Transformer (GPT), and was introduced by OpenAI (Radford et al., 2018). This is the architecture used for the models discussed throughout this dissertation. The basic transformer architecture was first introduced by Vaswani et al. (2017), and featured both an encoder and a decoder model. At the heart of

the transformer’s success is the attention mechanism (Parikh et al., 2016; Kim et al., 2017), which allows the model to determine the importance of each step in a sequence, modelling token dependencies beyond their positions. This mitigates one of the significant issues with RNNs, allowing the model to carry information over longer sequences and enabling parallel processing. The description of the decoder-only transformer below is simple on purpose, omitting architecture choices such as different attention mechanisms, transformer block design, or activation functions, etc. Since this dissertation is not about building or training LLMs, this section describes only the forward pass to generate sequence continuations. The transformer architecture is not designed to explicitly model any aspects of creativity: there is no single component or function in that architecture that is responsible for behaviours specifically targeting creativity. I refer the reader to an overview of formal algorithms for transformers by Vaswani et al. (2017) and Phuong and Hutter (2022) for further details on LLM architectures and training algorithms that are not covered in this dissertation.

Architecture

The transformer architecture (see Figure 2.1) mainly consists of a stack of transformer blocks. These transformer blocks can vary in design, but for example, GPT-2 (Radford et al., 2019) uses the following design: multi-head attention layer, followed by a layer normalisation (Ba, Kiros & Hinton, 2016), followed by a fully-connected layer, then another layer normalisation. The multi-head attention (Equation 2.5 and Section 2.2.1 for the detailed description) and the fully-connected layer have residual connections (Deng et al., 2009) to mitigate the vanishing gradient problem (Hochreiter, 1998) and stabilise training for these deep networks. In this section, I first discuss how the input is encoded, followed by the attention mechanism and the multi-head attention layer, and finally the output layer.

Input Encoding The transformer maps an input sequence, a tokenised piece of text, $\mathbf{x} = (x_0, \dots, x_t)$, to an output sequence $\hat{\mathbf{x}} = (x_1, \dots, x_t, x_{t+1})$, where t is a time step. The decoder-only model is trained autoregressively, meaning it predicts the next token x_{t+1}

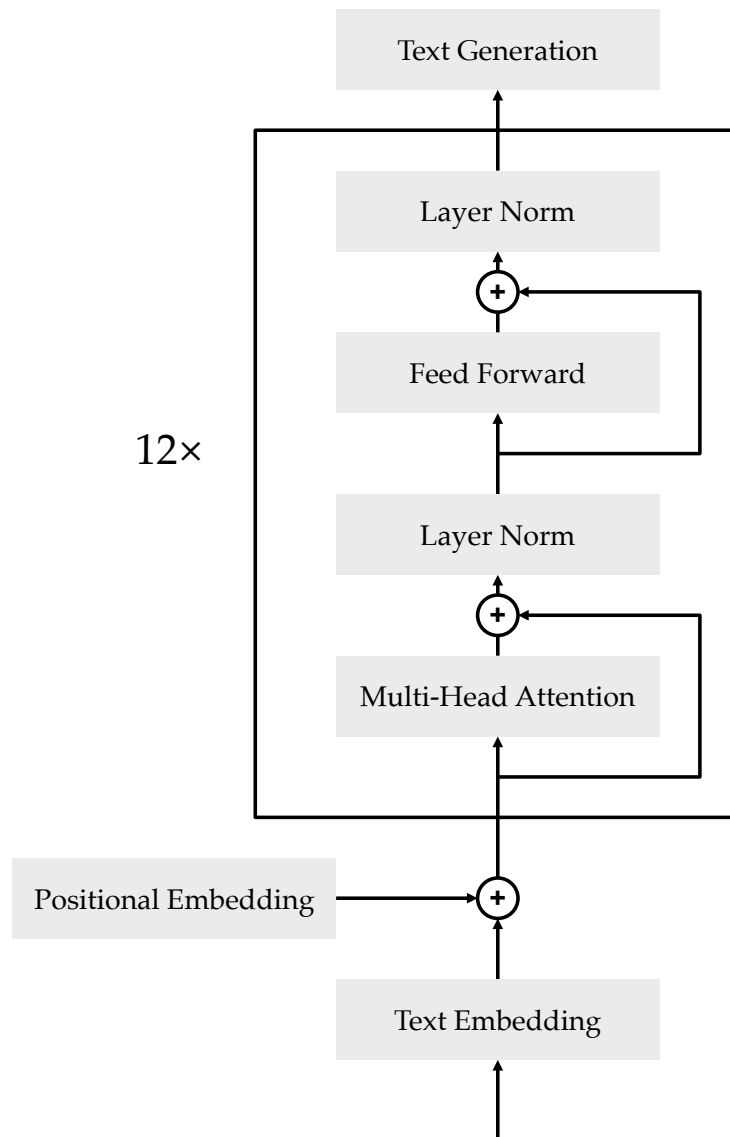


Figure 2.1: The decoder-only transformer architecture used for GPT-2. This is an example of the architecture used for the models discussed in this dissertation.

on any given input. Effectively, the objective of the model is to predict the input shifted right. If the input length is shorter than the maximum input length of the transformer, the input is usually left-padded with padding tokens or end-of-sequence (EOS) tokens.

The input is first encoded using a learned embedding matrix $W_e \in \mathbb{R}^{d \times v}$ and a learned positional embedding matrix¹ $W_p \in \mathbb{R}^{d \times l}$, where d is the embedding dimension, v is the vocabulary size, and l is the maximum input length. The positional embedding is necessary since the transformer does not use convolution or recurrence to distinguish tokens in different positions of the sequence. W_e and W_p are used as lookup tables, hence the embedding $\mathbf{e} \in \mathbb{R}^{l \times d}$ of input sequence $\mathbf{x}$ is:

$$\mathbf{e} = W_e[:, x[t]] + W_p[:, t]. \quad (2.1)$$

As such, the initial input to the first transformer block is a sequence of l embeddings, one for every step in sequence $\mathbf{x}$.

Scaled Dot Product Attention Attention is the key mechanism that enables transformers to make use of context and learn dependencies over sequences. The (single-query) attention function (Equation 2.2) works by computing the dot product of the query $\mathbf{q}$ and keys $\mathbf{k}$ vectors, scaled by a factor² of $1/\sqrt{d}$, before applying the softmax function to obtain the attention weights. The scaling is necessary as the magnitude of the dot product grows large as the dimension d increases (Vaswani et al., 2017). Finally, the attention weights are applied to the values $\mathbf{v}$ to get the weighted attention values, which are passed on to the next layer.

$$\text{SingleQueryAttention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}}\right) \mathbf{v} \quad (2.2)$$

where the softmax function is a function that normalises a vector $\mathbf{z} \in \mathbb{R}^n$ into a probability distribution (i.e. each element is in the interval $(0, 1)$ and all elements sum to 1):

¹The original transformer by Vaswani et al. (2017) uses positional encoding instead of W_p .

²The scaling factor is the average amount by which the length of a vector increases as d increases (see <https://peterbloem.nl/blog/transformers>).

$$\text{softmax}(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_j^n \exp(z_j)} \quad (2.3)$$

However, in practice, it is more efficient and easier to process multiple queries and combine them into matrices Q , K , and V :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{S}{\sqrt{d}}\right) V, \text{ where } S = QK^\top \quad (2.4)$$

For unidirectional decoder-only transformers, it is important to mask attention weights to prevent the model from peeking into the “future”. Otherwise, it would “see” the token it is predicting. Each row in the score matrix S represents a step of the input sequence $\mathbf{x}$; as such, we can mask S by setting the values in its upper triangle to $-\infty$.

Multi-Head Attention Analogous to words having different meanings in different contexts, multiple attention “heads” allow the transformer to learn different transformations for the same input. If we were to use only a single attention transformation, the layer would only learn one perspective.

A multi-head attention layer (Equation 2.5) consists of H heads. Each head consists of three linear layers, which apply a linear transformation to the inputs X and Z using $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{s \times l}$ for $i = 1 \dots H$, where $s = d/H$, yielding matrices Q , K , and V , respectively. After concatenating the attention-weighted values for each head, they are mapped to the output values using a linear layer $W^O \in \mathbb{R}^{Hd \times l}$. In the case of a decoder-only transformer $Z = X$, giving rise to self-attention. An encoder-decoder would keep Z , the encoded representation of the initial input sequence, fixed throughout the generation process, while X is the input of the previous generation step. This is known as cross-attention.

$$\begin{aligned} \text{MultiHeadAttention}(X, Z) &= \text{concat}(\text{head}_1, \dots, \text{head}_H) W^O \\ \text{where } \text{head}_i &= \text{Attention}(XW_i^Q, ZW_i^K, ZW_i^V) \end{aligned} \quad (2.5)$$

Model Output The last transformer block yields the final encoding h of sequence $\mathbf{x}$. Depending on the task the network is modelling, the last part of the network is specific to the task at hand. For example, sentiment analysis needs a configuration that classifies the input as positive, neutral, or negative. For the text generation task, the final layer is a fully connected layer that maps h into the outputs $\hat{\mathbf{y}}$, then the softmax function is applied to yield the next-token probability distributions $\hat{\mathbf{p}}$.

$$\hat{\mathbf{p}} = \text{softmax}(W_u^\top h) \quad (2.6)$$

where $W_u \in \mathbb{R}^{d \times v}$ is the unembedding matrix³. To generate a continuation from the original input, the next token is sampled from the last probability distribution $\hat{\mathbf{p}}_l$. The selected token is then concatenated to the input sequence $\mathbf{x}$ and fed back into the transformer, and so on, until it generates the EOS token or reaches some predetermined maximum number of steps.

There are various strategies for selecting the next token. For example, selecting a token using random sampling or taking the one with the highest probability. However, it turns out these simple sampling methods are prone to generating unnatural-seeming language. In the next section, I will discuss both sampling and search-based decoding strategies and how they can make the LLM behave more naturally.

2.2.2 Decoding Strategies

Decoding strategies for LLMs generally come in two flavours: sampling- and search-based. The key difference between search and sampling is that the former considers

³This matrix is sometimes referenced as the *softmax matrix*, since it is the transformation applied before the softmax, or simply *embedding matrix*.

multiple candidates while the latter considers only one. Both methods are repeated until the EOS token is generated or until it reaches some predefined limit.

Sampling-based Strategies The simplest form of sampling is greedy sampling, where the process at each time step simply selects the token with the highest probability given an input from $\hat{\mathbf{p}}_l$ (Wiher, Meister & Cotterell, 2022; Welleck et al., 2024). While this yields deterministic outputs, they are not necessarily of high quality, fluency, or coherence. Alternatively, we can use ancestral sampling (Wiher, Meister & Cotterell, 2022; Welleck et al., 2024), picking a token randomly from $\hat{\mathbf{p}}_l$ to introduce variety. Temperature is a common hyperparameter used with ancestral sampling. This modification is sometimes called temperature sampling (e.g., Renze, 2024). Temperature controls the shape of the distribution as high values flatten the distribution, while low values sharpen it (see Chapter 4). Ancestral sampling can still easily produce incoherent output (Welleck et al., 2024), especially if combined with high temperature values: it increases the chance of selecting the unlikely tokens in the tail of the distribution.

The reason GPT-2 (Radford et al., 2019) was successful is largely due to top- k sampling (Fan, Lewis & Dauphin, 2018). This decoding strategy truncates the next-token probability distribution, leaving only the first k tokens for sampling, where k is a hyperparameter. It turns out, this kind of truncation is essential for LLMs to produce fluent and coherent language, and mitigates a problem known as the softmax bottleneck (Yang et al., 2018). In human language, the vast majority of tokens most of the time have zero probability, for example, there are only so many continuations for the token sequence “sub, st, an”. This phenomenon occurs because the rank of the final layer is bounded by the size of the hidden state, which is much smaller than the vocabulary, the model cannot represent the high-rank distribution natural language requires; its practical consequence is that a non-zero probability is assigned to infeasible tokens. However, the softmax function always assigns some non-zero probability, causing an overestimation of token probabilities in the tail of the distribution. As such, truncation strategies are more accurate approximations of the true distribution (Hewitt, Manning & Liang, 2022; Finlayson et al., 2024).

There are various alternatives to top- k that all truncate the distribution according to different criteria. A common alternative is nucleus sampling (Holtzman et al., 2020), also known as top- p sampling, which selects the tokens of a sorted probability distribution in decreasing order until they cover the cumulative probability mass defined by hyperparameter $p \in (0, 1)$. Two other methods, ϵ - and its variation η -sampling (Hewitt, Manning & Liang, 2022) use a probability threshold. The valid tokens using ϵ -sampling are all tokens where $P(x_t | x_{<t}) > \epsilon$. η -sampling uses an entropy-dependent threshold, such that $P(x_t | x_{<t}) > \eta$, where $\eta = \min(\epsilon, \alpha \exp(-H(X | x_{<t})))$ and hyperparameter $\alpha = \sqrt{\epsilon}$.

Similarly, locally typical sampling (Meister et al., 2023) limits the valid tokens to those within a certain absolute range from the conditional entropy (expected information content) for a model at time step t ; in other words, the token should not be too predictable, which can lead to dull and repetitive output, or surprising, which can be incoherent or degenerate. First, $P(X | x_{<t})$ is sorted in increasing order according to $|H(X | x_{<t}) - \log P(x_t | x_{<t})|$. Then, the set of valid tokens is the set of tokens that cumulatively cover τ probability mass, similar to nucleus sampling. While there are other approaches to truncation, these are the most common and efficient approaches to truncating the distribution and mitigating the softmax bottleneck problem.

Not all sampling methods focus on truncation to better approximate the true distribution of the language. The Mirostat algorithm (Basu et al., 2021) generates a sequence of a predetermined average surprise (perplexity). It works by adapting the top- k hyperparameter according to the level of surprise, leaving more room if the perplexity is lower than the target and vice versa if the perplexity is higher. Methods such as top- k and nucleus sampling can degenerate and fall into a “boredom or confusion trap” while Mirostat avoids these over longer sequences (Basu et al., 2021).

Each of these approaches has its own underlying motivation. For example, η -sampling aims to generate more coherent and fluent text in long sequences; locally typical sampling is inspired by how humans use language; and nucleus sampling is shown to give a similar diversity of outputs as text written by humans (Holtzman et al., 2020).

Search-based Strategies The purpose of search-based methods is to use beam search to find interesting sequences (usually sequences with high probability) that would have been missed by sampling methods. Beam search is a classic heuristic that uses a pruned variant of breadth-first search, but it can be adapted to other search algorithms, such as A* (Hart, Nilsson & Raphael, 1968) or best-first search (Meister, Vieira & Cotterell, 2020). Beam search tracks a fixed number of candidates and ranks them according to an objective function. Since search is about finding candidates that are not readily found using sampling-based strategies, these methods often involve some notion of community diversity. There are various functions to encode diversity, for example, Self-BLEU (Zhu et al., 2018) or distinct n -grams (e.g., Li et al., 2016; Meister et al., 2023). But, in general, a diverse set of objects is as internally dissimilar as possible. Search increases output diversity and quality, but has a higher computational cost compared to sampling, since search algorithms generate multiple candidates and then select the best one according to some criteria (Welleck et al., 2024).

Naturally, some tasks require a different objective that does not focus on the probabilities of the output according to the model. For example, Contrastive Decoding (Li et al., 2023) uses beam search where a larger model (expert) is penalised by a smaller model (amateur) to prevent the expert from following bad language patterns learned by the amateur. The authors show that Contrastive Decoding leads to more diverse outputs from the expert (as measured by the average distinct n -grams counts), than other search-based methods, but is less effective than nucleus sampling.

For generating diverse sequences, an effective search-based method is Diverse Beam Search (DBS; Vijayakumar et al., 2018). Again, diversity is about the dissimilarity found in a set of objects; as such, DBS considers a fixed number of groups of beams in the generation process, each group lagging one time step behind the previous group. The objective function combines the joint probability of beams and the dissimilarity (using a variety of similarity functions to model different notions of diversity) between groups of beams at each step, resulting in more diverse outputs compared to standard beam search.

Franceschelli and Musolesi (2024a) introduced Creative Beam Search (CBS), a variant of DBS that focuses specifically on creative behaviour. They use the LLM as an evaluation function (LLM-as-a-judge) for response validation. The LLM generates a diverse set of responses, which are then reranked according to the LLM's evaluation. The validation steps ask the LLM to choose, from four candidates, which one it finds most creative (assuming that it can accurately judge the creativity of the output). CBS scores slightly higher than DBS when asked which one is more creative in a direct comparison task. The authors suggest that this can improve DBS, but a key point to keep in mind is that diversity and creativity are not synonymous. While diversity is a key component of creativity (Jordanous & Keller, 2016), the participants in the human study were only asked about creativity: they might have valued a different dimension of creativity.

There has also been some success using Monte Carlo Tree Search (MCTS), which is commonly used for training RL-based agents and was pivotal in the success of AlphaGo (Silver et al., 2016). Specifically for machine creativity, MCTS has been applied as a decoding strategy for recipe generation (Taneja, Segal & Goodwin, 2023) using GPT-2 (Radford et al., 2019). Taneja, Segal and Goodwin (2023) introduced a four-stage reward function to find interesting recipes; the generated recipes proved indistinguishable from human-created ones.

The benefit of search as a strategy for candidates is that you can define a preliminary objective function that filters candidates, which is a clear downside of sampling. Each method mentioned above has its own conceptual basis, and they each aim to increase the range of potential outputs with the aim of finding interesting samples to then choose a single exemplification. This is especially helpful for creative tasks. In this dissertation, I do not explore search-based methods, but in Chapter 5, I present a sampling decoding strategy that could be reformulated as a search objective.

2.2.3 Instruct Models and Preference Optimisation

Fine-tuning has been very effective at making pre-trained LLMs more useful for many (creative) applications, such as poetry generation in the style of specific authors (Sawicki et al., 2022, 2023b). However, a key innovation to make LLMs generally applicable is by tuning them for instruction following, as this allows them to be flexible and perform many tasks (Ouyang et al., 2022), and simplifies their user interface.

Instruction-tuning was initially done using Reinforcement Learning with Human Feedback (RLHF; Christiano et al., 2017). The process starts with supervised fine-tuning of the base model on a dataset for a downstream task, such as translation, summarisation, conversation, and so on. The goal of this step is to train the model to deal with data and prompt formatting as required for instruction following (Figure 2.2). Supervised fine-tuning is followed by a preference tuning phase that uses an actor-critic algorithm, such as Proximal Policy Optimisation (Schulman et al., 2017), to further optimise towards human preference. The general idea is to use a policy model that serves as the reward function to improve the LLM to align with human preferences. This policy model is trained on sampled data from the LLM annotated by humans as either positive or negative. A significant downside of the actor-critic algorithm is stability, and it requires very high-quality data. RLHF is largely superseded by Direct Preference Optimisation (DPO), a technique proposed by Rafailov et al. (2023). This technique avoids the RL paradigm and its downsides while maintaining the same objective by reverting to a classification scenario. As with RLHF, DPO uses a positive-negative pair of samples from the LLM and aims to assign high probability to the positive sample and low probability to the negative sample. DPO has some limitations: it is less effective if the supervised fine-tuning (SFT) step yielded a poor model (Xu et al., 2024a), and it seems to generalise less effectively compared to explicit reward modelling (Lin et al., 2024).

Recently, a technique called Reinforcement Learning with Verifiable Rewards (Lambert et al., 2025) was introduced. It introduces the final RL-step, after SFT and DPO, that uses

```
<|system|>
{{system_message}}
<|user|>
{{instruction}}
<|assistant|>
{{response}}<|endoftext|>
<|user|>
{{instruction}}
<|assistant|>\n
```

Figure 2.2: Instruct LLMs are trained on data that follows a specific templating pattern to allow conversation. This figure shows the template used for OLMo 2, but other models use similar formatting.

a scoring function that verifies the output, such as Math-Verify⁴, another LLM-as-a-judge, or simply executing code, instead of the human preference policy model. This technique is used in the OLMo 2 family of models (Team OLMo, 2025). Similarly, Xin et al. (2024) uses an interactive theorem prover as a reward model. The goal of these methods is not necessarily to optimise preference, but to put emphasis on requirements for a certain task. For example, on factuality and accuracy in mathematical reasoning or generating code.

All these methods push the LLM to a more constrained space of possible outputs, which has significant implications for creative tasks, as the LLM can repeatedly generate very similar outputs (see chapter 4). In chapter 5, I investigate this particular issue by looking at the diversity in LLM outputs as these fine-tuning steps are applied.

⁴A system for parsing and evaluating mathematical expressions in text. See <https://github.com/huggingface/Math-Verify>

2.3 Narrative Generation: Traditional Approaches and LLMs

This dissertation is *not* focused on building creative systems that generate narratives, but instead we use this task to test the research questions about LLM behaviours. As such, in this section, I do not extensively survey narrative generation, but aim to draw several analogies and insights between programs designed for narrative generation and the LLMs useful to this dissertation. For a detailed perspective on narrative generation and CC, I refer to a survey by Gervás (2009). Narrative generation has been a grand challenge of machine creativity going as far back as the 1970s when early programs to generate narratives, such as Novel Writer (Klein et al., 1973) and TALE-SPIN (Meehan, 1977a), were created. Over the past decades, there have been many projects that perform narrative generation. Hence, it is an interesting creative domain for the research questions this dissertation aims to tackle.

2.3.1 Traditional Narrative Generation Systems

Programs that generate natural language and narratives have been an active research area for decades. Early examples of narrative generation programs are Novel Writer (Klein et al., 1973), TALE-SPIN (Meehan, 1977a), TAILOR (Smith & Witten, 1991b), and MINSTREL (Turner, 1994). These symbolic programs use predetermined templates that are filled according to various rules and associations in a knowledge base, which is also why their abilities are more limited in terms of fluency. The key mechanism for these systems is their process and how they plan the story, checking if the narrative is logically correct and consistent. Since then, there have been symbolic systems, such as Fabulist (Riedl & Young, 2006) and MEXICA (Pérez y Pérez & Sharples, 2001) that attempt to go beyond the limitations of the older systems by expanding their world model and introducing self-reflection, respectively.

More recently, Markov chains provide a way to generate narratives that are more akin to LLMs since they generalise from training data and sample from the resulting model.

Simple low-order Markov models generally perform poorly and suffer from data sparsity (as mentioned in section 2.2), but have found some success using constraints and heuristics guiding generation (Ammanabrolu et al., 2019a; van Stegeren & Theune, 2019; Harrison, Purdy & Riedl, 2021). Naturally, LLMs have replaced Markov models and found their way into many modern narrative generation systems (e.g., Yuan et al., 2022; Kelly et al., 2023; Méndez & Gervás, 2023; Ye et al., 2023; Góngora et al., 2024).

2.3.2 Narrative Generation using Large Language Models

Unlike traditional systems, pure LLMs do not have the ability to plan a story and must rely heavily on sampling methods. Nonetheless, there are some parallels between traditional systems and LLMs. The example stories generated by TALE-SPIN in Meehan (1977a) are, besides language fluency, surprisingly similar to what a Llama 2 70B Chat model generated using simple prompts (See appendix B.2) due to existing patterns and standard elements like “Once upon a time there was...”. The Grandmother program (Casebourne, 1996) uses an interactive process for the user to provide additional knowledge, not unlike the idea of in-context learning for LLMs (Brown et al., 2020). In-context learning also allows one to mimic the process of iteratively improving on the previous output by providing hints, as demonstrated by Agirrezabal and Gonalo Oliveira (2025). This is akin to the engagement-reflection loop (Aguilar & Pérez y Pérez, 2015), a part of MEXICA (Pérez y Pérez & Sharples, 2001) that allows it to iteratively improve its own outputs. The interaction of a user with the LLM can be seen as a co-creative game, which yields similar output compared to human-human collaboration (Vidra et al., 2025).

By contrast, a pure LLM is more limited, since it does not explicitly model creative processes, such as planning a story, as researchers have done for traditional narrative generation systems. This does not mean an LLM cannot implicitly plan a story, it just means we do not know to what extent the LLM applies such planning. The more useful application of LLMs for narrative generation is using them to increase fluency and flow, and generate narratives that read less like a list of actions and more like actual stories. But in this dissertation, the focus is on questions about LLM behaviour, rather than

questions about the creative process of LLMs performing the narrative generation task.

2.4 Towards an Understanding of LLM Creativity

This chapter presents the broader background supporting this dissertation. Specifically, I discussed a view on creativity, creativity evaluation, LLMs and the transformer, decoding strategies for LLMs, and finally, traditional approaches to narrative generation and how they connect with LLMs. Each of the remaining chapters has a dedicated background section expanding on the relevant literature. Importantly, in the next chapter, I will further discuss the application and use of LLMs for creative tasks in the context of traditional concerns about AI, embodiment and the LLM as a creative evaluator.

Chapter 3

Large Language Models and Computational Creativity Evaluation Frameworks

This chapter is based on: Peeperkorn, M., Brown, D. G., & Jordanous, A. (2023). On characterizations of large language models and creativity evaluation. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 143–147). Association for Computational Creativity

3.1 Introduction

Popular awareness of Large Language Models (LLMs) rose abruptly with the November 2022 release of ChatGPT and the later 2023 release of GPT-4 (OpenAI, 2023), which allowed anyone to interact with an LLM. This led to headlines about issues ranging from high-school essays and the future of news columnists' jobs to a massive influx of generated stories submitted to the sci-fi/fantasy magazine *Clarksesworld* (Acovino, Kelly

& Abdullah, 2023). This raises questions about the creative abilities of LLMs, which has large societal implications: not just for increasing creative productivity, but, in a more pessimistic view, for displacing human creative work (Lamb, Brown & Grossman, 2025) or for flooding the world with AI slop, leading towards a cultural collapse (Heuser, 2025). There are legitimate concerns that we are collectively worse off when individually performing creative tasks using AI assistance (Doshi & Hauser, 2024). Hence, it is paramount to understand the creative abilities (or limitations) of LLMs and generative AI more broadly.

The Computational Creativity (CC) community has put decades of effort into developing methodologies, criteria, and evaluation frameworks for assessing the creativity of machines. LLMs, and generative AI in general, appear to outperform ‘traditional’ CC systems that use a systems approach to generating creative artefacts (Morain, Kinghorn & Ventura, 2023). Yet, the generative process driving transformers is comparatively simple: repeatedly drawing a next token from a learned probability distribution. This raises the question: Can we evaluate the creativity of pure large language models using existing CC frameworks?

To address this, I examine pure LLMs (see Chapter 1) through three established CC evaluation frameworks: the Creative Tripod (Colton, 2008) which assesses *perceived* creativity of a system, FACE (Colton, Charnley & Pease, 2011; Pease & Colton, 2011a) which characterises the *creative acts* a system can perform, and the Mere Generation framework (Ventura, 2016) which tests if a system has moved beyond the “*mere generation*” threshold. This chapter presents a domain-general systematic analysis rather than a domain-specific empirical study; therefore, I do not apply empirical evaluation frameworks such as Ritchie’s Empirical Criteria (Ritchie, 2007) (see Section 3.2.3) or SPECS (Jordanous, 2012).

The analysis in this chapter is the initial step towards the dissertation’s broader concern: the creative application of LLMs. How LLMs execute creative tasks depends on what we can say about their creative behaviour as systems, and the established CC frameworks examined here provide a first route into that question. As we shall see after the analysis, existing frameworks are not fit to establish the creativity of LLMs, and we need to resort

to empirical analysis. The main reason why the evaluated frameworks are not fit for establishing the creativity of LLMs is that we cannot point to components of the LLM responsible for a particular aspect of creativity that these frameworks define as necessary for creative behaviour.

3.2 Background

In this section, I extend the background provided in Chapter 2. First, I give a brief overview of transformer-based language models, how they are used, and how they are tuned to perform tasks; then I explore what creative tasks they might perform and how they are used as creativity evaluators; and finally, I discuss what creativity evaluation is in the context of CC.

3.2.1 Transformer-based Language Models

LLMs have dramatically changed both CC and natural language processing (NLP). A transformer-based language model (Vaswani et al., 2017) establishes a probability distribution over token outputs based on properties of the token distribution in the training data (see Chapter 2). In particular, at each position, the influence of previous tokens in both the prompt and the output of the LLM can vary: the LLM can maintain the name or gender of a character across a story and multiple outputs and can produce sentences with high fluency. The degree to which a language model can understand how much earlier tokens influence later tokens depends on the parameters of the model and how it processes sequences. Transformers with billions of parameters are better at maintaining long-distance dependencies, because they allow sequences to be trained in a parallel fashion, as opposed to the sequential signal propagation of recurrent neural networks, where the signal can vanish quickly (Hochreiter, 1998). This allows for far more subtle interactions between the tokens of a paragraph over longer text (Vaswani et al., 2017).

Besides model size, the volumes of training data have similarly exploded, enabling models to work with an astronomical variety of information. As a result, the probability distributions can be implicitly *conditioned* by prompt engineering: one can alter the type of response obtained by changing the rhetorical tone of the prompt, e.g., “I bet you don’t know the answer to this question:”, or by giving a role or a persona in the prompt, e.g., “You are a sceptical scientist: do vampires exist?” (Khatun & Brown, 2023). The ability to invoke new modes or personas (Kojima et al., 2022) allows style changes of the model (Sawicki et al., 2023a) in both obvious ways: “You are Walt Whitman; write a poem about a clam.”, and less obvious ways: “You hate poetry and think it’s a waste of time; please write a review of this poem:”.

Transformers can be enhanced in a variety of ways. First, it is necessary to make a clear distinction between instruct models complemented with RLHF (Christiano et al., 2017; Ouyang et al., 2022) and similar approaches (see Chapter 2, Section 2.2.3), and those that are not. While instruction-tuned models are much more useful, these tweaks may come at a cost: models could outright refuse to write violent fiction or even discuss important political speeches or religious texts (Ganguli et al., 2022; Perez et al., 2022), or instruction-tuning could reduce output diversity (see Chapter 5). This potentially puts a substantial dent in their creative potential.

A further development in this line is that current LLMs are trained to perform intermediate token generation, which is sometimes referred to as reasoning or thinking (Kambhampati et al., 2026). It works by introducing a special token during fine-tuning that marks the start of an “internal dialogue” or “reasoning trace”, which extends the prefix the model conditions on before it commits to a response (DeepSeek-AI et al., 2025a). I will refer to the tokens produced in this phase as the *intermediate token trace*, following Kambhampati et al. (2026) in avoiding the labels “reasoning” and “thinking” for what remains a sequence of generated tokens. While tweaking for human preference might limit the creative potential, intermediate token generation may allow the model to explore niche solutions, potentially finding novelty.

An alternative frame for changing the overall distributions of LLMs is to alter their

probability distributions by fine-tuning on a specific corpus of data as opposed to human preference data: the model keeps its fluency while generating sentences consistent with the shape of the training data. In addition to generating poetry in a particular author's style (Sawicki et al., 2022), this approach can also yield transformers that do better at answering basic mathematical problems or making valid logical arguments (Cobbe et al., 2021).

3.2.2 LLMs and Creative Tasks

LLMs appear to be excellent systems for combinational and exploratory creativity (Boden, 1992), endlessly synthesising new combinations and variations, in the hope of finding a creative artefact. Prompting a model for variations of an idea can simulate creative brainstorming. Indeed, one delightful application is to use them in a Mad-Libs style, to fill in holes in sentences in surprising ways (e.g., Garimella et al., 2020), exploring low-probability words in the transformer's conditional distribution.

One astonishing feature of the LLM is its ability to imitate the style of authors, which is a genuine creative task on its own (Brown & Jordanous, 2022). It can easily, given enough training data, rewrite a few sentences in another style, including a style not attached to an individual, such as "the style of a fourth grader". Such prompts allow the attention mechanism to "shift" the probability distribution to vocabulary words and sentence structures used by these simulated personas (Jiang et al., 2023b). LLMs can combine styles of poets or authors and interpolate between the two, enabling one to explore, mix, and match between different styles, stories, and other ideas (Sawicki et al., 2022, 2023b). The personas do not have to be just the style of known authors (Sawicki et al., 2023a); the LLM can spit out variations of the same content endlessly, using prompts like, "rewrite this sentence as though it has been written by a fourth-grade student", or "summarise this argument in 100 words or fewer".

In CC, LLMs are useful not only for generating creative artefacts, serving as poem and story generators or other writing assistance tools, but also for evaluating creative output.

However, this requires understanding and knowledge, especially if these evaluations are then put directly into the world. Consider an LLM that is asked to evaluate jokes (Goes et al., 2022). It is prompted with a joke (the object) and various personality descriptions (conditions), and asked if it is “funny” or “not funny”. Testing a joke against multiple personalities then allows exploring how the joke works for different people and backgrounds, or at least how the LLM predicts it would work. LLM-as-a-Judge has become common practice to supplement evaluation of tasks that do not have verifiable criteria, such as many creative tasks (Zheng et al., 2023). and it appears to work reasonably well as exemplified by Sawicki et al. (2025) showing strong correlations with human judgements.

3.2.3 Creativity Evaluation

Evaluation is an essential part of CC research; we must understand the abilities and limitations of the creative systems that we build, as already discussed in Chapter 1. Over the years, there have been multiple proposals towards this goal, often focusing on different criteria. Generally, evaluation is done by judging the system by its artefacts, often without the observer even knowing that the artefact was created by a computer program, in order to avoid the risk of bias (Colton, 2008). However, in practice, this is difficult; for example, if a work generated by a machine is deemed creative by competing in an art gallery, it usually requires full disclosure of the origins of the work. Nonetheless, this idea of artefact-based evaluation is based on the principle that when we celebrate the excellence of an artefact, we are, in fact, celebrating the process that brought it into being and not just the creator of the product itself (Ritchie, 2007; Colton & Wiggins, 2012). Like Boden (1992), Ritchie focuses on the out-of-context artefact as a proxy for the creative process: if an object is of value, its production must have been as well. Over the past decades of creativity research, a standard definition emerged that relies on two characteristics: an artefact must be *novel*, and it must be of *value* (Stein, 1953; Barron, 1955; Runco & Jaeger, 2012). Novelty concerns whether the artefact is original and new, and value if it is useful, of high quality, or otherwise worth something to someone. Building on the standard characteristics, Ritchie (2007) was the first to present an artefact-based evaluation framework using empirical criteria, adding a *typicality*

characteristic to assess whether an artefact is an example of the artefact type in question, arguing that otherwise evaluation is nonsensical. However, Ritchie’s empirical criteria are increasingly difficult to apply to large models like LLMs. The main reason is that the empirical criteria require an inspiring set against which to evaluate typicality and novelty. For generative AI models, the inspiring set (the training data) is very large and unknown¹, making Ritchie’s empirical criteria impossible to use. In this chapter, the focus is not yet on artefact-based evaluation, but instead on three frameworks that I will introduce in the next section: how algorithms use and represent knowledge (Ventura, 2016), the perception of creativity by external reviewers using three notions of skill, appreciation, and imagination (Colton, 2008), and how creative a system can be, by describing what creative acts it can perform (Colton, Charnley & Pease, 2011). These frameworks approach an evaluation by examining the system itself and the process it uses to determine whether a system should be deemed creative.

3.3 Computational Creativity Evaluation Frameworks

Here, I introduce three established CC evaluation frameworks used for the analysis presented in this chapter: “Mere Generation” (Ventura, 2016), the Creative Tripod (Colton, 2008), and FACE/IDEA (Colton, Charnley & Pease, 2011; Pease & Colton, 2011a). Each of these frameworks has its own approach to assessment and its own criteria for creativity, reflecting its evaluation goals.

3.3.1 Mere Generation

Systems that “merely” generate are viewed negatively by some in the CC community (Ventura, 2016), and are claimed by some to be unable to exhibit creative behaviours (Veale, 2015). Some object to this characterisation and argue that this kind of blind variation is precisely what is driving the creative process in humans (Simonton, 1999,

¹Usually described as “a new mix of data from publicly available sources” (Touvron et al., 2023b).

2023). To clarify what “mere generation” means, Ventura (2016) originated a thought experiment, presenting a spectrum of seven main prototypical algorithms² to determine when we cross the “mere generation” threshold into more creative systems. The complete spectrum is available in Appendix A.1.

To evaluate each algorithm, Ventura uses the two standard criteria and introduces a third: intentionality. First, the two standard criteria, novelty and value, are defined by Ventura (2016) as follows:

Definition 2 (Novelty). *The quality of being new, original or unusual; this is relative to the population of artifacts in the domain in question and can apply in the personal or historical sense.*

Definition 3 (Value). *The importance, worth or usefulness of something; this would typically be ascribed by practitioners of the domain in question.*

Intentionality complements the two standard characteristics as a way to determine that the system must have had some role in creating the artefact, and that it is not simply the result of a completely random process, which additionally requires the system to have a way to represent knowledge. To acknowledge this requirement, the third criterion, *intentionality*, is defined as follows:

Definition 4 (Intentionality). *The fact of being deliberative or purposive; that is, the output of the system is the result of the system having a goal or objective—the system’s product is correlated with its process.*

Ventura’s spectrum itself (Table 3.1) starts with random sampling and increasingly introduces more knowledge to the system. Initially, this involves plagiarising and memorising an inspiring set, which would be considered neither novel nor valuable according to the definitions above. The next algorithm is about generalisation, building a model based on the inspiring set, finding patterns to extrapolate from in order to produce novel outputs. This culminates in algorithms that evaluate their own work using their perceptual abilities, such as filtering bad outputs, using knowledge bases to inform generation, and thereby conveying meaning intentionally.

²There are two additional algorithms which are excluded because they are specific to haiku generation.

Table 3.1: The spectrum of seven prototype generation algorithms. Appendix A.1 reproduces the complete algorithms as defined by Ventura (2016).

#	Process	Description
1	Randomisation	Generation using a stochastic process.
2	Plagiarisation	Generation by plagiarising an inspiring set I .
3	Memorisation	Generation by memorising an inspiring set I .
4	Generalisation	Generation by generalising an inspiring set I .
5	Filtration	Generation by modelling an inspiring set I and filtering candidate artefacts via a fitness function $\text{fitness}()$.
6	Inception	Knowledge-based generation by modelling an inspiring set I , employing a fitness function $\text{fitness}()$ and leveraging a knowledge base K .
7	Creation	Creative generation by modelling an inspiring set I and leveraging a knowledge base K followed by evaluating the perception of the generated artefact.

To determine if a system has moved beyond “mere generation”, Ventura gives the following five candidate thresholds that could be fulfilled, each with increasing demands:

- (1) it can be demonstrated to possess any knowledge whatsoever
- (2) it can be demonstrated to possess knowledge that it has had some hand in structuring/acquiring
- (3) it can be demonstrated that it has some reasonable chance of producing both novelty and value
- (4) it can be demonstrated that it has some reasonable chance of producing both novelty and value and at least one of these is intentional
- (5) it can be demonstrated that it has some reasonable chance of producing both novelty and value and both of these are intentional

These exact thresholds are still up for debate, but Ventura (2016) argues that CC as a field has moved beyond “mere generation”: its systems satisfy some or all of the thresholds. CC systems have indeed become very sophisticated at using and remixing knowledge

to produce novel and valuable artefacts. However, the current class of state-of-the-art generative models appears capable, yet their underlying processes are perhaps “merely generative”, just sampling from a distribution. Yet, they readily produce more complete artefacts (e.g., Ramesh et al., 2022; Agostinelli et al., 2023; Sawicki et al., 2023a) than some of the more intricate CC systems not based on huge training datasets (e.g., Pérez y Pérez & Sharples, 2001; Ventura, 2019; Bodily & Ventura, 2024a). The framework has also guided system building: Bodily and Ventura (2024a) designed Pop*, a computational system for music composition that operationalises some of framework’s requirements for knowledge representation and intentionality.

3.3.2 The Creative Tripod

The Creative Tripod framework (Colton, 2008) is centred around the analysis of external *perception* of creativity in CC systems. The framework is heavily inspired by visual creativity, and painting in particular, but can be applied generally. For example, Colton (2008) applies it to the HR system (Colton, 2002) and the Painting Fool (Colton, 2012), which forms mathematical theories and conjectures. The motivation for using the perception of creative behaviour is to create a level playing field for machine and human creativity, to eliminate or reduce bias. Colton (2008) argues that each painter, automated or otherwise, should ultimately develop their own style; thus, arguing that it is necessary to omit quality from the criteria for analysis of creativity.

The Creative Tripod considers two factors in its evaluation process: it assesses creativity based on the artefact of the system, and it examines the processes that the system executes from an external perspective. The framework consists of three legs (hence a “tripod”): skill, appreciation, and imagination, defined as follows:

Definition 5 (Skill). *A system is skilful if the program is highly capable of creating new artefacts. Every creative system needs some capacity to produce something.*

Definition 6 (Appreciation). *A system is appreciative if it has some form of measure by which it can evaluate its own artefacts.*

Definition 7 (Imagination). *A system is imaginative if it has the capacity of a program to come up with novel, potentially surprising outcomes; to do something that goes beyond the programmer's initial intention or expectations.*

The program needs to fulfil all three main criteria for it to be deemed creative. Each leg is split up into three parties: the software, the programmer, and the consumer. If the software satisfies the three criteria, then, regardless of the computer or the programmer, it must be considered creative. However, the more parties contribute to perceived skill, appreciation, and imagination, the more creativity we should project onto the system.

Although Colton tries to avoid a notion of quality, the assessment of artefacts inevitably leads to some form of value, which is linked with quality (Runco & Jaeger, 2012). However, an important distinction with other frameworks is that it is about the system's ability to assess itself or its artefacts.

A key issue for this framework, as Colton (2008) mentions, is that the bias against machine creativity is significant and puts the artificially creative system at a disadvantage. Generative AI is increasingly integrated into creative workflows (Garcia, 2024) and appears to influence human creativity favourably (Zhou & Lee, 2024). The experience of using generative AI can vary for different reasons, such as resources or specific artistic circumstances (Kaila, Holzapfel & Jääskeläinen, 2024). While these more recent findings are about collaboration with these models and not about their perceived creativity, an interesting open question, which would greatly impact the Creative Tripod, is whether public opinion is shifting; the negative bias against machine creativity or the issues with authenticity (Colton, Pease & Saunders, 2018) might be less strong, as artificially generated artefacts are now more prevalent in society. However, an issue with perception might be that, given the quality of artefacts that LLMs produce, external perception becomes a weak basis for assessing quality, as it might be too easily satisfied.

Table 3.2: The eight individual creative acts of the FACE model as presented in Pease and Colton (2011a).

Act	Description
F^p	a method for generating framing information
F^g	an item of framing information for an A , C , or E act, generative or process
A^p	a method for generating aesthetic measures
A^g	an aesthetic measure for process or product
C^p	a method for generating concepts
C^g	a concept
E^p	a method for generating expressions of a concept
E^g	an expression of a concept

3.3.3 FACE/IDEA

Colton, Charnley and Pease (2011) initiated what they call Computational Creativity Theory (CCT), which currently consists of the FACE and IDEA descriptive models. In the spirit of Computational Learning Theory (Angluin, 1992), CCT is a formalism that aids in studying the design and analysis of creative systems and assessing progress. The FACE/IDEA models are inspired by ideas from creativity research in psychology and by the analysis of human creative acts (Pease & Colton, 2011a). While CCT seems to be competing with another formalism, the Creative Systems Framework (Wiggins, 2006a, 2006b), the purpose is different; the focus here is not on formally describing how creativity works, but rather on the impact, development, and comparison of creative systems. For example, Colton et al. (2015) use FACE to assess whether a society can accept and integrate such creative systems. Another example of a FACE evaluation is by Cook, Colton and Pease (2012) on the ANGELINA system, an automated game designer.

The purpose of the FACE model is to describe and assess the creative acts a system can perform to compare it with other creative systems. This set of acts is expressed as a tuple of creative acts F , A , C , and E , where F is an item of framing information, A is an aesthetic measure, C is a concept (an executable program, or something that can be interpreted as such), and E is an expression of a concept (Table 3.2). An aesthetic measure is the

application of a function that takes as input a pair of creative acts (C^g, E^g) (either can be null if the system is unable to perform that creative act) and returns a positive score. Thus, a creative act is defined as follows:

Definition 8 (Creative Act). *A creative act is a tuple of the individual acts a system performs in producing an artefact, drawn from framing information F , aesthetic measure A , concept C , and expression E , where each element is marked g if it yields an artefact of that type and p if it yields a method for producing such artefacts.*

The FACE model not only considers creative acts that produce artefacts, but also those that create new processes. For example, a system might “mix” components of existing evaluation measures to compose new ones (e.g., Linkola & Hantula, 2018a). The distinction between these two kinds of creative acts is that one produces artefacts, while the other produces new functionality for the system. As such, each of the four types of creative acts has a “generative” and “process” variant denoted by a g and p superscript, respectively. Hence, FACE distinguishes a total of eight creative acts (Table 3.2), with a clear distinction between “make something” and “make something that makes something”. The tuple description of a creative system is a subset of the eight individual creative acts. Since FACE allows a system to create new processes, it is important to make a distinction between a creative act of the system and one performed by a third party, for example, a human evaluator. This kind of creative act is denoted by a bar, for example, $\overline{A}^g$ is a user-performed aesthetic measure of generated artefacts. Lowercase characters are used to describe a specific element of a creative act, for example, a^g denotes a specific aesthetic function of the system.

In addition to describing the creative acts a system can perform, Colton, Charnley and Pease (2011) argue that the FACE model is useful to quantitatively and qualitatively compare creative systems. For the *quantitative comparison*, we can define an ordering of descriptions to determine which system is more or less creative than another. For example, if we agree that a system creating aesthetic functions is more creative than a system that only applies aesthetic functions, we can describe this as follows: $\langle A^p \rangle > \langle A^g \rangle$. These orderings are useful to describe desiderata for assessing progress of the creative

abilities of CC systems (Colton et al., 2015). The *qualitative comparison* evaluates creative acts based on the outcomes of the aesthetic functions. For example, a user-defined aesthetic function $\overline{a^g}$ can be used to evaluate a set of concept/expression pairs $S = \{(c_1^g, e_1^g), \dots, (c_n^g, e_n^g)\}$ generated by the system, and then calculate an average aesthetic score: $1/|S| \sum_{i=1}^{|S|} \overline{a^g}(c_i^g, e_i^g)$.

By contrast, the IDEA model focuses on the impact of creative acts on the actors (the programmer, the user, and the audience) who can be involved in the development of the creative system. The model is named after its Iterative Development-Execution-Appreciation cycle and identifies six engineering stages of the creative system, each describing the extent of the impact of the creative acts undertaken by the creative system. The impact of the stages ranges from what we already know, to something novel, to performing creative acts that produce artefacts too novel to be judged meaningfully. The audience plays a key role in the IDEA model, as it provides the ratings, well-being, and cognitive effort used in the measures defined by Colton, Charnley and Pease (2011) to assess the development process of a creative system. The analysis presented in this chapter does not use IDEA, as training and building LLMs is not within the scope of this dissertation.

A particularly relevant question is how appropriate the FACE framework is for state-of-the-art generative AI. One complication is that they are machine learning systems, and Colton, Charnley and Pease (2011) describe these as $\langle C^g \rangle$, and yet LLMs appear to do more once trained.

3.4 Creativity Analysis

LLMs exhibit substantial new capabilities and behaviours that warrant an evaluation of their creativity. This raises the question of whether LLMs satisfy the CC frameworks described above. Specifically, I examine whether LLMs have moved beyond “mere generation” (Ventura, 2016), whether they can be perceived as creative (Colton, 2008),

and what creative acts they can perform according to the FACE descriptive framework (Colton, Charnley & Pease, 2011; Pease & Colton, 2011a).

3.4.1 Scope: The Creative System

The creative system that is the subject of this analysis is the pure LLM as defined in Chapter 1 (Definition 1), with its architecture described in Chapter 2. That is, the analysis concerns LLMs without any additional tooling, augmentation, or interaction. The goal is not to establish whether an LLM is creative, but to determine whether the pure LLM satisfies the frameworks. If the pure LLM satisfies the frameworks, then any superset (the pure LLM with any additional tool, augmentation, or interaction) will only enhance its creative capabilities. The definition of a pure LLM does not exclude training data, and the key distinction between then and now lies in instruction tuning and, in particular, in training for intermediate token generation (DeepSeek-AI et al., 2025a). Such changes to the training data still fall within the definition of the pure LLM and arguably yield considerably more capable models: hence, I specifically consider instruction-tuned LLMs in this analysis.

To summarise, the system under consideration is a text-only *instruction-tuned* transformer-based language model that takes as input a sequence of tokens, which are translated into embedding vectors. These vectors are processed by the transformer blocks, which incorporate the attention mechanism (Section 2.2), and by a final unembedding layer, to yield a probability distribution that is modified by a decoding strategy (Section 2.2.2) and from which the next token is drawn. The token is then appended to the input, and the process repeats to produce the next token, until the model generates an end-of-sequence token, upon which the output is returned to the user.

3.4.2 Methodology

As discussed above, each framework has its own objective, criteria, and approach. The steps taken in the analysis for each framework are as follows:

- *Mere Generation* (Ventura, 2016) requires the following steps:
 1. Triangulate the position of the system on the spectrum (Table 3.1) based on its architecture and the criteria.
 2. Determine which thresholds have been met (Section 3.3.1).
 3. If the system is located on the spectrum and has some role in producing the output, then the thresholds determine how far it has moved beyond mere generation.
- *The Creative Tripod* (Colton, 2008) requires the following steps:
 1. For each leg (skill, appreciation, and imagination) and each party (the program, the programmer, and the consumer), determine what each party contributes to the perception of creativity.
 2. If the *program* meets the criteria for each leg, it is perceived as creative; the results for the programmer and the consumer may lend further weight to this conclusion.
- *FACE* (Colton, Charnley & Pease, 2011) requires the following steps:
 1. Describe the system in terms of the creative acts (F , A , C , and E , and their variants g and p) that it can perform through its architecture.
 2. The creativity of the system is determined by (the size of) the tuple resulting from the analysis.

3.4.3 Analysis

Mere Generation

Are LLMs beyond “Mere Generation”? I now work through the spectrum to determine where the LLM is situated. LLMs draw their next tokens from a non-uniform distribution and have thereby moved beyond Algorithm 1. They also readily satisfy Algorithms 2 and 3: they can be fine-tuned to engage in plagiarism or memorise inputs by overfitting

on a smaller dataset (Sawicki et al., 2022), and there are documented cases in which they reproduce news articles near-verbatim (Grynbaum & Mac, 2023). However, their well-documented tendency to hallucinate (e.g., Kalai et al., 2025) and to fabricate outputs indicates that they do not merely retrieve an artefact from a database, as Algorithms 2 and 3 describe; their behaviour exceeds retrieval. This brings the analysis to Algorithm 4. From an architectural perspective, this is where the LLM is situated: it is a learned function that generalises over its training data in order to generate new artefacts. The subsequent algorithms all introduce a loop that produces artefacts and evaluates them until a criterion is satisfied. From a purely autoregressive perspective, the LLM is incapable of performing such a loop in order to iterate on its complete outputs and make revisions. It does appear to have some capability to reflect internally, as Lindsey et al. (2025) found that LLMs do plan ahead in some way, for example, when generating a rhyme. This means that in their parameters and embeddings they have captured some way of meeting the necessary criterion, e.g., filtering out words that do not rhyme.

The intermediate token trace used by recent models extends this further, as it can be seen as a proxy for the components required by Algorithms 5 and 6: the fitness function, and the exploration of the knowledge base within the LLM, which, via the attention mechanism, contributes to the final output.

While the trace is merely a proxy, and cannot be seen as concrete evidence, it can be argued that LLMs acquire *referential* grounding (whether their representations are about anything in the world at all) through the process of fine-tuning rather than from the corpus they are trained on (Mollo & Millière, 2023). Next-token prediction over a corpus does not by itself secure reference; a base model holds a great deal of information and knowledge but is not yet a competent language user, meaning it is unable to link its utterances to the phenomena they concern. What fine-tuning for human preferences adds is a feedback loop in which outputs are selected for their success in use, and it is this relation to the world, rather than the corpus, that lets the model use its knowledge base successfully.

Algorithm 7 of the spectrum requires perceptual and sensory capabilities, and since

this analysis considers a unimodal text-based LLM, it does not satisfy this algorithm. However, through becoming a competent language user, the LLM can refer to, discuss, and potentially apply phenomena beyond its own sensory reach (Kouwenhoven, 2025). This is similar to the finding that sighted and blind people share an understanding of colour that is enabled by linguistic communication (Kim et al., 2021). This suggests that LLMs can perform Algorithm 7, at least to some extent, without explicit sensory machinery.

Nevertheless, LLMs exhibit behaviours suggesting that some aesthetic function, or some application of a knowledge base, is present internally, in the form of the learned embedding space and the weights of the model. In prompting an LLM, a degree of responsibility and intentionality (Definition 4) is evidently deferred to the model. A minimal prompt instructing the model to write a story yields a wide range of valid and distinct outcomes (as seen in Chapter 4).

Result The spectrum provided by Ventura is indexed to specific mechanisms. By mechanism, LLMs are placed at Algorithm 4, and yet, through their intermediate token generation capability, they exhibit behaviours associated with Algorithm 7, which gives two verdicts. This renders Ventura’s framework *undecidable* in the case of LLMs. LLMs evidently meet threshold 5 (Section 3.3.1) and have a reasonable chance of producing output that is both novel and valuable, and both may be regarded as intentional, in the sense that intermediate token generation lets the model elaborate on a problem and reflect on its approach before committing to an output, thereby arriving at an interesting result. However, the evaluation criteria provided by Ventura (2016) remain difficult to apply, given how LLMs represent and use “knowledge”. The intermediate token trace provides only a proxy; it is not a true fitness function or knowledge base available for inspection. Consequently, it remains difficult to trust an LLM’s responses when questioning it for evaluation and, more importantly, to determine what its constraints are and what intention truly lies behind the response.

Table 3.3: What each party contributes to the perception of creativity for each leg of the Creative Tripod. The program is split by whether it is fine-tuned for intermediate token generation.

Party	Skill	Appreciation	Imagination
Program (no intermediate tokens)	using the attention mechanism and truncation strategies,	via the attention mechanism over the prefix	hallucinations, search-based decoding strategies
Program (intermediate tokens)	the program yields fluent, coherent human-like text	+ the intermediate token trace might be a proxy for the appreciation function	+ the intermediate token trace
Programmer	data curation and training schemas		
Consumer	the prompt: the task specification	the prompt: criteria the output must satisfy	the prompt: themes and styles for the artefact

The Creative Tripod

Can LLMs be perceived as creative? The Creative Tripod’s objective is satisfied if the program is skilful, appreciative, and imaginative, so I start there; Table 3.3 summarises what each party contributes to each leg. The key mechanisms that give the LLM its skill are the attention mechanism, which maintains dependencies and consistency over the sequence, and a decoding strategy that truncates the probability distribution. For example, top- k sampling allows a transformer language model to generate coherent and fluent human-like text (Fan, Lewis & Dauphin, 2018), and nucleus sampling (Holtzman et al., 2020) has been shown to make generated text as diverse as human-authored text. Both methods counteract the overestimation of probabilities for tokens in the tail of the distribution, which stems from the softmax bottleneck (discussed in more detail in Chapter 2). The autoregressive loop not only allows it to generate something, which is the basic requirement, but, in combination with the attention mechanisms and truncating decoding strategies, allows it generate high quality outputs.

Colton (2008) appeals to the processes within the system to argue for its ability to appre-

ciate and imagine. Pure LLMs run an autoregressive loop, generating token by token, which means that the attention mechanism doubles as a form of appreciation, but only over the current prefix (the initial input and the generated output so far). This is appreciation in a limited sense: without further extensions to the system, the model cannot judge a full draft and revise it before presenting the final output to the user.

Imagination is usually achieved through a search algorithm (Colton, 2008), so an LLM that employs a search-based decoding strategy (Section 2.2.2) can be considered to imagine. An LLM that employs a sampling-only strategy is more limited, but I would argue that hallucination is still an imaginative behaviour, as it may yield an unexpected yet felicitous outcome when applied to, for example, poetry. This works primarily because the consumer is able to internally make interesting connections; where an output is required to be factually correct, hallucinations are of course undesirable.

Regarding the consumer and the programmer, the former has a modest role in increasing the perception of appreciation and imagination by supplying such information in the prompt, for example, criteria the output must satisfy, or themes and styles for narratives, while the latter plays a key role in enabling the general creative potential of the LLM through data curation, architecture design, the choice of decoding strategy, and the training regime (including fine-tuning steps). This is not just limited to skill, but can also influence appreciation and imagination, for example by choosing training examples in the style of a specific author, or of a specific tone.

Fine-tuning is hugely influential on an LLM's performance, in particular, for models that are trained for intermediate token generation (e.g. DeepSeek-AI et al., 2025a); hence, it is important to make this distinction in this analysis. For systems that do *not* produce intermediate tokens, we are unable to identify an explicit appreciation function, or a description of such a function, even though the output makes it reasonable to believe that one exists. For a model that is trained to perform intermediate token generation, the intermediate token trace might be viewed as a proxy for the appreciation measure the LLM applied to the output, although there are no guarantees. I would argue that the intermediate token trace is a direct example of imaginative behaviour, because it allows

the model to search, by means of the attention mechanism, for relevant sequences that should form part of the final output. It should be stressed that this is a weaker claim than the one Kambhampati et al. (2026) reject: I do not take the trace to be reasoning, nor to be a faithful account of how the output was produced, but only to be an inspectable record of the search that the frameworks under evaluation here demand we be able to point at.

Result As a framework, the Creative Tripod does not demand much of a system for it to be considered creative, and the general architecture of the LLM, together with a decoding strategy, readily satisfies its criteria. However, none of these mechanisms were made part of the system with creative behaviour in mind, and it remains impossible to inspect the system and point to a particular appreciation function that would strengthen the case for creativity, as could be done for traditional creative systems such as HR, whose suite of interestingness measures (Colton & Muggleton, 2006) supports its creative behaviour. For the LLM the case is less clear: the intermediate token trace might be a fair proxy, but it is heavily reliant on the programmer’s data curation and training setup.

FACE

What creative acts can LLMs perform according to the FACE model? Table 3.4 summarises the analysis that follows. Let us start with the concept C , which relates to the model itself. The model requires architecture design choices, data curation, and a training schema, all decided by the programmer. This yields $\overline{C^g}$ under the FACE formalism, with the overline indicating that this act is provided by the programmer and not by the program itself. At this moment, the LLM cannot perform C^p , as it would require an end-to-end pipeline taking over the choices made by the programmer, and thus a significant extension of the system, which is not within the scope of this analysis. Once we have a trained LLM, it can be used to generate expressions E^g ; different expressions can be generated as each token is selected through the attention mechanism and the decoding strategy (which are parts of $\overline{C^g}$), and as the process continues it effectively creates a new generation function, thus allowing the LLM to perform E^p . To evaluate the framing and aesthetic functions

of the FACE formalism, we again need to distinguish between LLMs with and without intermediate token generation. Without intermediate tokens, we cannot determine if the LLM performs A^g , because we cannot point to a concrete function; at best we can argue that such a function is performed implicitly and opaquely, but this would not match how the FACE formalism is defined by Colton, Charnley and Pease (2011). Similar to the argument made for the Creative Tripod, the intermediate token trace is potentially a reasonable proxy, but not concrete evidence, for a family of aesthetic functions that weigh in on the final output. Finally, and interestingly, the intermediate token trace can be inspected, which could provide a piece of framing information for the final output that contributes to the perceived creativity of the system, meaning an LLM is capable of the F^g creative act when it performs intermediate token generation. The p variants of A and F are not performed in either case, because that would require updating the model itself to change either how it generates intermediate tokens or the implicit aesthetic function it applies.

Result If we disallow intermediate token generation, then the LLM is described as $\langle E^g, E^p \rangle$; if we allow it, then $\langle F^g, E^g, E^p \rangle$. If we significantly relax the requirements and allow implicit aesthetic functions (which would not satisfy the aesthetic function as defined in the FACE formalism), or view the intermediate token trace as a valid proxy, then A^g is also performed.

These are minimal tuples that omit all the related creative acts performed by the programmer or the user. Obviously, the model itself is the result of the programmer training it, meaning we could include $\overline{C^g}$. Additionally, we could include $\overline{C^p}$ as the programmer's choices about architecture and decoding strategy, and $\overline{F^g}$ or $\overline{A^g}$ in the prompt as another piece of framing information, or aesthetic measure, respectively. However, these are prerequisites rather than components that are required during the generative process, and are therefore omitted.

Table 3.4: The eight creative acts of Table 3.2 for a pure LLM as described in Section 3.4.1. A dash (–) means the creative act is not represented.

Act	Pure LLM	Pure LLM with Intermediate Token Generation
F^g	–	The intermediate token trace provides a piece of framing information towards the final output.
F^p	–	–
A^g	At best implicitly performed, but no function can be identified to determine what was performed and how	The intermediate token trace is a proxy for a family of aesthetic functions
A^p	–	–
C^g	A trained model itself, but it is reliant on architectural decisions and the training data curated by the programmer, yielding $\overline{C^g}$	
C^p	–	–
E^g	An output; the artefact the model produces	
E^p	Attention mechanism and decoding strategy	

3.5 Discussion

From the discourse so far, we can deduce that, although LLMs began as mere next-token generators, they now also exhibit behaviours that go beyond their computational mechanisms and are implicit in the parameters. This disparity causes trouble for the CC evaluation frameworks, as we are unable to use them to show how and why LLMs are creative; we can only show that they appear to satisfy their respective criteria.

The Mere Generation framework (Ventura, 2016) appears to be in the most precarious situation. On Ventura's spectrum of seven prototype algorithms, it has become clearer that LLMs sit, from a computational perspective, at Algorithm 4: create an object by building a model using an inspiring set (training data) and then generalising from it. On the other hand, some outputs of LLMs are easily at the top end of the spectrum as they *appear* to handle semantics very well. Although I mentioned that LLMs do have a reasonable chance of crossing all five lines in the sand provided by Ventura (2016) and therefore go beyond "mere generation", the issue is that LLMs find themselves in two locations at once on the spectrum: LLMs behave intentionally, or not? As such, I conclude that the "mere generation" framework is no longer useful to evaluate CC systems if they involve generative AI models with complex knowledge representations.

Similarly, the FACE model (Colton, Charnley & Pease, 2011) wants us to define certain creative acts the LLM seems able to perform, such as applying an aesthetic function, but we are unable to point to such a function. This makes it impossible to claim that the LLM can perform this act, since we do not know how or what it is actually doing, yet the output gives the appearance that some measure is (somehow) applied. One way to argue that LLMs can use aesthetic measures could be to include a description of the measure in the prompt (Sawicki et al., 2025). However, it remains unclear how the LLM then applies this aesthetic measure, and this kind of steering is overall not very reliable (Bowman, 2023), assuming this still applies to the current state-of-the-art. The Creative Tripod (Colton, 2008) faces similar issues regarding appreciation, as we do not know how these models use and apply it when provided with instructions. While LLMs readily

satisfy the Creative Tripod, its bar for creativity is quite low, as one could argue that the system just needs a generator function, a search algorithm, e.g. A* (Hart, Nilsson & Raphael, 1968), and some evaluation function, e.g. data compression (Cilibrasi, Vitányi & de Wolf, 2004), to fulfil its criteria.

One framework the analysis leaves standing is the Standardised Procedure for Evaluating Creative Systems (SPECS) presented by Jordanous (2012). SPECS is a “meta” framework: it aims to standardise the creativity evaluation methodology and does not take a specific perspective on creativity or define what creativity means. The procedure itself is deliberately simple: identify a definition of creativity appropriate to the system and its domain, state the standards that follow from it, then test the system against those standards and report the results. Each of the three frameworks examined above breaks at the point where it asks something about the inside of the system: the Creative Tripod needs us to locate skill, appreciation, and imagination; FACE needs us to point at an aesthetic function; “mere generation” needs us to locate where knowledge and intention reside. SPECS asks none of these questions, but instead standardises how the researcher sets the criteria: it prescribes how to evaluate but not what should count as creativity, so it has to be filled with criteria drawn from somewhere else. Although SPECS does not dictate that the evaluation should be empirical, as a “meta” framework it allows for such an evaluation, whereas the others do not. As such, it has informed the evaluation approach in this dissertation without being formally applied.

From a broader CC perspective, it appears that “bitter lesson” (Sutton, 2019) applies to creative systems as well. Sutton observed that general methods, such as neural networks, are ultimately more effective than those that integrate human knowledge, provided they can scale computation and utilise the available data. CC has been all about designing creative processes and building models by defining rules and aesthetic functions, but the bitter lesson tells us that this is likely not as effective as providing vast amounts of data to a general-purpose model to generalise from (Algorithm 4). In this sense, choosing the pragmatist’s approach (van Dijk et al., 2023) that argues it is too easy to dismiss any notion of understanding, over the stochastic parrot perspective (Bender et al., 2021), is perhaps the only stance to move forward. This will have a significant impact on the way

CC research is conducted, as it needs to rely more on empirical evaluation and requires methodology standardisation for assessing the desired properties.

A very recent paper³ by Ventura (2026) conducts a similar but more general evaluation of generative AI, covering both FACE and the Mere Generation framework. While the conclusions are largely similar, there are some distinct differences. My conclusion is that LLMs are architecturally at Generalisation (Algorithm 4) and that, by means of intermediate token generation, they can implicitly perform the remaining algorithms on the spectrum, while Ventura (2026) argues that it is not clear whether generative AI systems can go beyond Filtration (Algorithm 5), and specifically argues that they cannot yet reach Creation (Algorithm 7) due to a lack of grounding. However, I argue that through referential grounding (Mollo & Millière, 2023), they achieve some of the necessary behaviours even when operating in a text-only modality.

The key difference between the FACE evaluation in this Chapter and (Ventura, 2026) is that the latter does not cover an LLM trained to perform intermediate token generation, which enables additional creative acts as per the evaluation presented in this chapter. This is an important distinction, as it is the result of choosing the training data, which Ventura (2026) considers part of the tuple in the form of $\overline{C\mathcal{S}}$. These two evaluations thus highlight differences in perspective on generative AI and, in particular, on whether we allow implicit or explicit evidence to determine if systems are creative.

3.6 Conclusion

Evaluating pure LLMs is hard: they are black boxes and learn complex patterns, and yet have a simple generation process. On an input–output level they appear to pass the test, but it is unclear *how* they produce an artefact, so these frameworks cannot meaningfully argue that they are creative. In particular, it is difficult to point at the mechanisms at the system level that are responsible for the processes required by “mere

³Presented at ICCV’26 in June 2026.

generation”, the Creative Tripod, and FACE/IDEA to consider which machine is creative and which is not. Unsurprisingly, LLMs can be considered creative systems by the standards of each framework: the behaviours of the LLM are not merely generative, despite its relatively simple architecture; there is a mechanism to argue for each leg of the tripod; and the LLM is able to perform creative acts on its own (not just user-provided creative acts). However, the frameworks are proving hard to apply due to the simplicity of the architecture, dependencies in training data, and reliance on implicit functionality. Crucially, the frameworks do not inform us exactly how the LLM is being creative, since we cannot explicitly locate responsible functions in the code. In conclusion, we have to rely on empirical methods to identify such functions, to understand what makes them creative, how they perform creative tasks, and more broadly, how the consumer can employ LLMs towards creative purposes.

Chapter 4

Is Temperature the Universal Creativity Parameter?

This chapter is based on: Peeperkorn, M., Kouwenhoven, T., Brown, D. G., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models? In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 226–235). Association for Computational Creativity

4.1 Introduction

Large language models (LLMs), in particular instruction-tuned variants, like ChatGPT and Claude, have become ubiquitous tools in society. Unsurprisingly, they have also found use for performing and assisting in creative tasks, such as writing stories (Calderwood et al., 2020), poems (Sawicki et al., 2023a), jokes (Veale, 2024), dialogues in video games (Volum et al., 2022), and more. For such creative tasks, *temperature* is often described as the parameter that enables creative behaviour in a language model (Manjavacas et al., 2017; Roemmele & Gordon, 2018; Chen & Ding, 2023). Temperature controls the

uncertainty or randomness in the generation process, leading to more diverse outcomes by levelling probabilities for candidate words. However, randomness alone does not capture a complex phenomenon such as creativity (Simonton, 1999, 2023); this would imply that noise is the most interesting thing one can create (Zenil, Delahaye & Gauchere, 2012). Creativity is a quality that involves multiple dimensions, such as social interactions and communication, or independence and freedom (Jordanous, 2016), that are not easily quantifiable. Yet, one probe into LLM creativity has found a limited positive effect on creativity at higher temperatures on a divergent association task (Chen & Ding, 2023); however, they do not account for the influence of other factors, such as different prompts. Temperature may have an effect on creativity, simply because without variation, nothing new can be created, but it is unclear if LLMs have sufficient “knowledge” so that randomness is all you need, which would be the case if temperature is the creativity parameter.

Following the analysis in Chapter 3, I present the first empirical investigation of this dissertation, which investigates how temperature affects the creativity of stories. LLMs are challenging to reliably probe, and most analyses of their capabilities and behaviours involve large benchmarks, which are challenging to interpret due to their scale and aggregation of outcomes. For example, many small LLMs appear to have difficulty with understanding multiple-choice questions (Khatun & Brown, 2024), yet these tasks are often included to assess the performance of LLMs in general. Reproducibility and reliability are further complicated by small differences in prompts that can lead to vastly different outcomes (Weber, Bruni & Hupkes, 2023). Moreover, the training data is often unknown, and we have no insight into the model architecture and additional plugins or processing for commercial LLMs. To improve reliability and reproducibility, I predetermine a *fixed* context: model and prompt. While a single model and prompt may not be sufficient to prove that temperature is the universal creativity parameter for any LLM, a single experiment can be enough to disprove this assumption. Key to the evaluation approach is the *exemplar object*, the greedy sample for that context, which is the baseline for our experiments. In this chapter, I present an empirical analysis and a creativity evaluation with human participants of stories generated by Llama 2 Chat

(Touvron et al., 2023b) on four necessary conditions for creativity: novelty, typicality, cohesion, and coherence. The results show that the influence of temperature on creativity is far more nuanced and weak than the “creativity parameter” claim suggests.

4.2 Background

The generative aspect of LLMs and other large generative models naturally sparked an interest for application towards creative tasks. The jump in output fidelity and ease of use is radically changing the landscape of creative tasks that a machine can perform. LLMs certainly seem capable of producing novelty, typicality, and grammatically fluent output (chapter 3), but it is difficult to determine how exactly to approach such a creativity evaluation (Baer & McKool, 2009; Lamb, Brown & Clarke, 2018). To investigate temperature as the creativity parameter, it is necessary to establish what is meant by creativity; what conditions are required for outputs of the LLM to satisfy?

4.2.1 Conditions for Creativity in Narrative Generation

As discussed in chapter 2, something is usually considered creative if it is novel or original and useful or effective (Boden, 1992; Runco & Jaeger, 2012). Usefulness is closely related to quality and value, as things of low quality tend not to be useful or have value. Because creativity is such a complex phenomenon, further properties have been proposed, such as surprise (Simonton, 2012; Grace & Maher, 2014) and typicality (Ritchie, 2007). Both of these properties are often viewed as variants of novelty, and the differences are very subtle. If something surprises you, there must be something that you have not seen before, something unexpected. Indeed, Martindale (1990a) draws a similar observation, noting that novelty is the disruption of expectedness. Typicality is more interesting because the difference is that before we ask any questions regarding novelty and quality, we must consider the prior question: to what extent is something typical? Is the object an example of its class or genre? Typicality and novelty are, like surprise and novelty, somewhat

at odds, as high typicality often implies low novelty; however, Ritchie (2007) suggests that typicality is a prerequisite for evaluating creativity, meaning that for something to be judged as novel it has to some extent be typical first. This view lends itself well to this work because we only consider the creativity of LLMs based on the artefacts it produces and not in a larger scope. While we do not use Ritchie’s empirical criteria directly, his perspective fits our purpose as we mainly evaluate the artefact of the system and not the process behind it, over a framework like the Creative Tripod (Colton, 2008), which specifically focuses on the behaviour of creative systems. Ritchie (2007) formulates the essential properties as follows: *novelty* is the extent to which the produced item is dissimilar to existing samples within its class, *typicality* is the extent to which the produced item is an example of the class in question, and *quality* is the extent to which the produced item is a high-quality example of its class.

For narrative writing tasks specifically, it is relatively poorly understood what is valued for creativity. For example, the importance of genre is unclear (D’Souza, 2021). In a large creative writing literature survey, D’Souza (2021) outlines various baseline criteria drawn from a large body of work. For example, a narrative needs to have an elaborate plot, an original voice, interesting characters, a good flow, linguistic originality, amongst many more. These criteria largely connect with the essential properties above, and are used to outline four necessary conditions to establish if temperature affects creativity for the narrative generation task in this study. Hence, for the task in this chapter, I require *novelty*, *typicality*, and *quality* in terms of *cohesion* and *coherence* (Graesser et al., 2004).

Novelty and typicality are relatively straightforward to reformulate in this domain. We would say that a story is *novel* if it is dissimilar to other stories, if it has some characteristics that set it apart. For example, it may be written in a novel setting or narrative style. A story is considered *typical* if it is of the same form as other stories in the same category. For example, based on stories we have read before, we can observe certain patterns, narrative style, and other characteristics. When we are presented with another story, we say it is typical if it follows similar rules or patterns and has those same characteristics.

Quality, in general, can mean a wide variety of things and can be defined depending on

what you value. For example, LLMs can easily and fluently write high-quality texts in terms of grammar and spelling, but this falls short for our purposes since the generated stories can be entirely meaningless or nonsensical. A key aspect of a good story is a good narrative that is cohesive and coherent. The difference between cohesion and coherence is subtle, as pointed out by Graesser et al. (2004). Cohesion is about the objective features in a text that hold everything together; in other words, sentences are well-connected, and characters, events, and actions are used appropriately. Coherence, however, is about the reader’s interpretation and the meaningfulness of the text. The reader should be able to easily follow and understand what it means. For our purposes, the quality of a story is a combination of cohesion and coherence: a story is considered *cohesive* if words and sentences are connected and grammatically consistent, and if characters, events and actions are used appropriately. A story is *coherent* if it is easy to follow the story and understand what it is about.

4.2.2 The Temperature Parameter

Temperature is a hyperparameter τ of stochastic generative models to regulate the randomness in a sampling process (Ackley, Hinton & Sejnowski, 1985). The softmax function (Equation 4.1) applies a non-linear transformation to the outputs of the final layer of the neural network, turning them into a probability distribution over possible outputs. The temperature parameter regulates the distribution’s shape, redistributing the output probability mass.

$$\text{softmax}(\mathbf{z})_i = \frac{\exp(z_i/\tau)}{\sum_j^n \exp(z_j/\tau)} \quad \text{where } \mathbf{z} \in \mathbb{R}^n \quad (4.1)$$

For $\tau > 1.0$, high probabilities are decreased, while low probabilities are increased, and vice versa for $\tau < 1.0$. Higher temperatures increase entropy and perplexity (Du, Yang & Welleck, 2025), leading to more randomness and uncertainty in the generative process. Typically, values for τ are in the range of $[0.0, 2.0]$, and $\tau = 0.0$, in practice, means greedy sampling, where one always takes the token with the highest probability.

4.2.3 Probing Large Language Models

There are roughly three techniques to “control” the LLM: through training data (either from scratch or fine-tuning), via in-context learning (Brown et al., 2020), and with different hyperparameter configurations. We focus here on using LLMs that are already instruction-tuned, meaning they are fine-tuned for conversation following a specific prompting format. An instruction-tuned LLM can respond to instruction and appear to “understand” what is asked of it (Brown et al., 2020). This suggests that by changing the context in the prompt, we shift the learned probability distribution accordingly. In a co-creative or interactive setting, one might ask to change details and refine previously generated poems or stories, allowing for more creative outcomes. In essence, by tweaking the prompt, we narrow down the valid range of the sampling distribution closer to the intended outcome. Adjusting the prompt to get the “right” slice of the probability distribution, however, is very sensitive to the choice of words and even punctuation (Weber, Bruni & Hupkes, 2023). It becomes increasingly difficult to separate the effect of the prompt as you allow more randomness via the temperature hyperparameter in the generation process.

If temperature is the creativity parameter, then one would expect something similar to what happens when one changes the prompt; that at higher temperatures, the outputs may appear novel because the process samples more broadly. However, it is difficult with this kind of investigation to reliably reproduce outputs and make comparisons across different models, prompts and hyperparameter settings, to draw any such conclusions. Even synonymous or highly similar instructions may yield very different outcomes as we show in Figure 4.1. In the next section, I will explain why the greedy sample is a suitable point of departure.

4.2.4 Prototypes and Exemplars

The method for evaluation we propose in this chapter is inspired by two theories of cognitive categorisation, namely prototype theory (Rosch, 1973) and exemplar theory

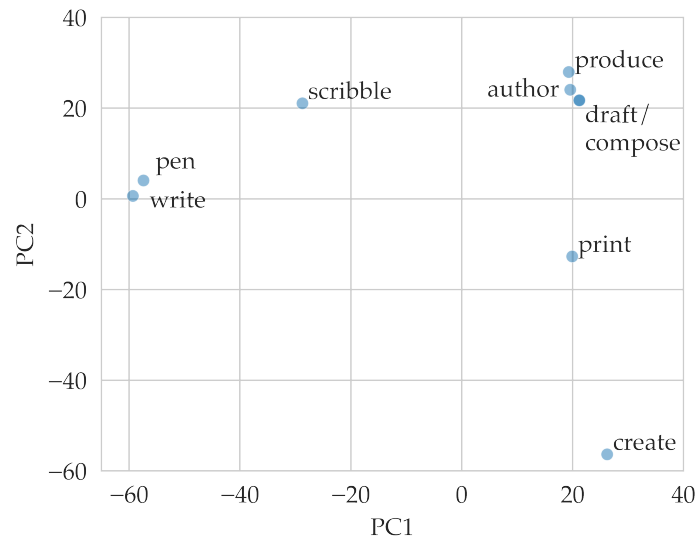


Figure 4.1: This figure shows a PCA projection of the Llama 2 Chat embeddings, where each point is a story written using Llama 2 Chat 70B with $\tau = 0$ (See section 4.3.2 for more details). The labels represent synonyms used for “Write” in the prompt; for example, “Write a story”, “Produce a story”, “Print a story” and so on. This indicates that evaluation is made more complex by the choice of words in the prompt, even though these verbs are semantically close. Interestingly, “Draft a story” and “Compose a story” yielded identical outputs.

(Medin & Schaffer, 1978). These two theories are often contrasted, but have much overlap. The key difference is that in exemplar theory, categorisation of new stimuli is not compared against a single prototype object, but against multiple exemplars of the category. A prototype is considered something that is derived from presented stimuli, where an exemplar is one of the previous stimuli. Prototype theory observes that many ideas and objects can be put in the same cognitive category based on their name, but as one looks closer, they may not have that many things in common. This relates strongly to the idea of *family resemblances* described by Wittgenstein 1953, where he argues that even though items may have similar names, there is not necessarily one thing common to all, but a complex network of similarities and relationships between members of the category. He uses the category of games to illustrate this observation:

Consider, for example, the proceedings that we call “games”. I mean board-games, card-games, ball-games, athletic games, and so on. What is common to them all? — Don’t say: “There *must* be something common, or they would not be called ‘games’ ” — but *look and see* whether there is anything common to all. — For if you look at them, you will not see something that is common to all, but similarities, relationships, and a whole series of them at that. To repeat: don’t think, but look! — Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball-games, much that is common is retained, but much is lost. — Are they all ‘*entertaining*’? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games, there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way; can see how similarities

crop up and disappear.

And the upshot of these considerations is: we see a complicated network of similarities overlapping and criss-crossing: similarities in the large and in the small.

—Wittgenstein (1953, §66)

In a category, basic objects (prototypes) are the ones that carry the most information, are most representative of their category, and their information can be used as a compressed representation of other class members. Rosch et al. (1976) show that basic objects are first categorisations made for that category of objects in an environment. This is reminiscent of exemplar theory, and people often use a mixture of prototypes and exemplars as they categorise objects. When people are primed with just exemplars, they can still accurately describe the prototype (Medin, Altom & Murphy, 1984).

These theories have further connection with conceptual spaces, geometric mental structures, where each point represents an object, the dimensions are properties, and concepts (categories) are convex regions (Gärdenfors, 2014). The central member, the prototype, or most typical object, has the most in common with most objects in the category, i.e. it has the shortest average distance to all other objects in the category. This does not mean that it has overlap with all other objects, but it has some resemblances with each member of the category. Conceptual spaces also play an important role in Boden's framework of creativity (Boden, 1992). She uses a metaphor, a map of the mind, to describe exploratory creativity and argues that novelty and value can be found through venturing into unknown regions. To be more precise, the typicality of an object relates to how well it fits with the exemplars or the prototype, while novelty could be viewed as how much the category changes or expands as a new stimulus is presented. If temperature were indeed the creativity parameter, then raising it should allow access to different regions of the output space. Prototype and exemplar theory provide a useful frame that motivates our approach to evaluation, and a method to evaluate further stimuli relative to a baseline (the central member or prototype). Next, I further explain what the exemplar is and how to generate it with LLMs in mind.

4.3 The Exemplar Story

Rigorously evaluating LLMs is hard. If we consider multiple models, it is difficult to compare their results because the learned representations and the training data could be very different. In the case of a single model, but many prompts, it is difficult to establish a solid baseline, and therefore also impossible to compare results, since lexical changes can have a cascading effect on the output. Hence, we need to work within a *fixed* context, using a specific model and prompt to establish a baseline that enables the comparison of results against independent variables, such as temperature.

4.3.1 Identifying the Exemplar

In section 4.2, I discuss why prototype/exemplar theory is a useful frame for a comparative evaluation. There are two perspectives on what we consider to be the prototype or the exemplars. First, since the peak of the LLM’s distribution condition on a prompt is the most probable or average outcome, using the prototype theory perspective, it is straightforward to view the greedy sample as the prototype. In that case, a category is every possible output with positive probability produced by one prompt and one model, with the greedy sample as its prototype. However, there are many prompts that can generate similar outcomes. The second perspective departs from this more abstract view. In that frame, the category is a group of prompts and a model, and the greedy object for each prompt is an exemplar. For the purposes of this work, I call the greedy output the *exemplar object*, but notably, prototype would also be a viable description, since there is only one prompt under consideration in the human evaluation experiment.

```
[INST]Write a story.[/INST]Here it is:\n\n
```

Prompt 4.3.1: The exact Llama 2 Chat prompt, showing the chat formatting and the suffix to make sure the model starts writing the story immediately.

4.3.2 Generating the Exemplar Story

We generate the exemplar and all other stories, using the instruction-tuned Llama 2 Chat 70B model (Touvron et al., 2023b). This instruction-tuned LLM is competitive with other models¹, but more importantly, it is open source², meaning that researchers have full access to the model architecture and weights. This is essential since the aim is to eliminate or minimise any effects besides the temperature parameter. Closed-source models, like GPT-3.5 and GPT-4, are for these reasons not viable as they are not transparent. Llama 2 is trained on an (unspecified) mix of publicly available data, and the chat variants are fine-tuned using reward modelling on annotated human preference data (Touvron et al., 2023b). For the instructions, I am using a minimal and neutral prompt (Prompt 4.3.1) so that it introduces as little additional information as possible. We applied 6-bit quantisation³, which reduces the model size with extremely minimal loss of quality (Zhang et al., 2023), to make it fit on available hardware. Llama 2 Chat 70B adds, unsurprisingly, much conversation to the output, instead of just performing the task, even when instructed to avoid this in the system prompt. This is unwanted behaviour, as it has a large effect on subsequent stories. To mitigate this, the LLM's supplied response starts with a minimal conversation to make sure that this is consistent across all generated stories, as in Prompt 4.3.1. Another benefit is that it reduced any post-processing for the human evaluation study, besides, occasionally, pruning one sentence at the end of the output that was not part of the story. We set top- k is set to 50, min- p is set to .05, the repetition penalty and top- p are both set to 1, meaning they have no effect, the seed is set to 4294967295. The truncation strategy (or a combination of strategies) is a potential confounding factor, as it truncates the probability distribution, meaning it is an upper limit on the number of tokens. For example, choosing different k in top- k , means that temperature redistributes the probability mass over a larger or smaller set of tokens. Noting that it will not change the order of the tokens. However,

¹At the time when this research was conducted in early 2024.

²The Llama 2 licence restricts commercial use, which strictly speaking is not open source as defined by the Open Source Initiative.

³Quantised using the llama.cpp library with setting Q6_K. See <https://github.com/ggerganov/llama.cpp/pull/1684>

choosing a truncation strategy is necessary to avoid degenerate outputs and overcome the softmax bottleneck (see Section 2.2.2 for more detail).

4.4 Case Study: “Write a story.”

The approach to investigating how temperature affects creativity in LLM-generated stories is a two-fold empirical analysis. First, I present a computational analysis to examine the behaviour of LLMs for different temperature values. In particular, I discuss how the stories are distributed in Llama 2 Chat’s embedding space and how to measure distance against the exemplar object. Secondly, I present an experimental study in which humans evaluate the creativity of generated stories to investigate the interaction between temperature and the four necessary conditions of creativity for a narrative generation task.

4.4.1 Computational Analysis

The goal of this analysis is to show whether temperature enables the LLM to access different regions in its embedding much further than those sampled at lower temperatures. We do not aim to evaluate stories for their creativity using the four necessary conditions, but establish any evidence from a computational perspective for the possibility of any exploratory creative behaviour, which Boden (1992) argues is essential for creative behaviour. Moreover, an increased diversity of outputs might suggest an increased likelihood of producing something novel. The computational analysis uses the setup described above, using Prompt 4.3.1 and its exemplar object. We generate 100 stories over 7 different temperature values $\tau \in [.001, .334, .667, 1.0, 1.334, 1.667, 2.0]$, to investigate the effect of temperature on the diversity of stories.

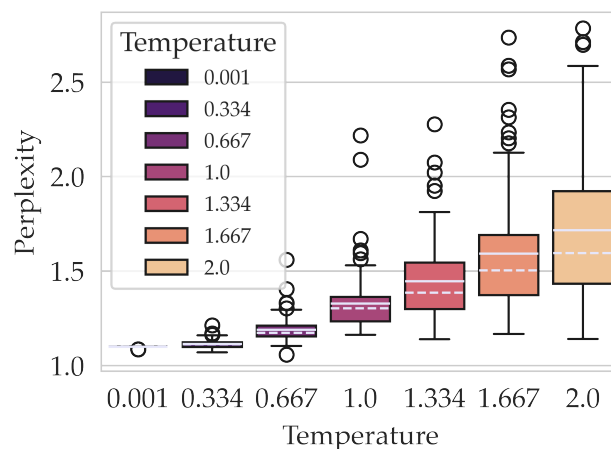


Figure 4.2: Perplexity of stories given the LLM per temperature value.

Evaluation Measures

One way to evaluate the quality of LLM-generated text is by using perplexity (e.g., Meister et al., 2023): higher perplexity implies more inarticulateness. This could be done with the same model. For purposes of this study, this does not inform much, as at higher temperatures the perplexity on average goes up (Figure 4.2), meaning that the quality goes down. Measuring quality using an independent model is more appropriate, but it raises the question of what is a suitable model for a narrative generation task. Moreover, previous work suggests that perplexity is a bad indicator of text quality compared to human evaluations (Pillutla et al., 2023; Zhang et al., 2023). For these reasons, we focus on diversity and similarity. An overall higher diversity proportional to temperature may indicate that different regions of the output space are used to generate the outputs. Comparisons for diversity and similarity can be done on the lexical level, which uses the texts directly, or on a semantic level by embedding stories in the LLM’s embedding space. Instead, the focus is on two commonly used measures that can be used to compare to the exemplar: on the semantic level, we use the cosine similarity between the embeddings of the stories and the exemplar, and for the lexical level, we use normalised edit distance (Levenshtein, 1966) between stories and the exemplar. Again, if temperature is a proxy

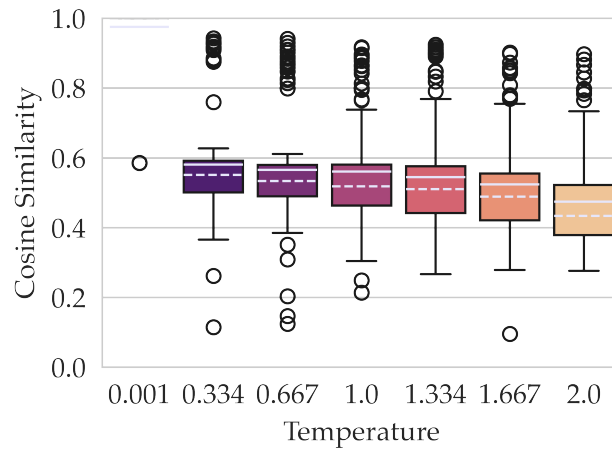
for creativity, then one would expect the outputs to extend into very different regions of the embedding space. We examine the distribution using principal component analysis (PCA), preserving global structure, to project embeddings of the generated stories into a two-dimensional space.

Results

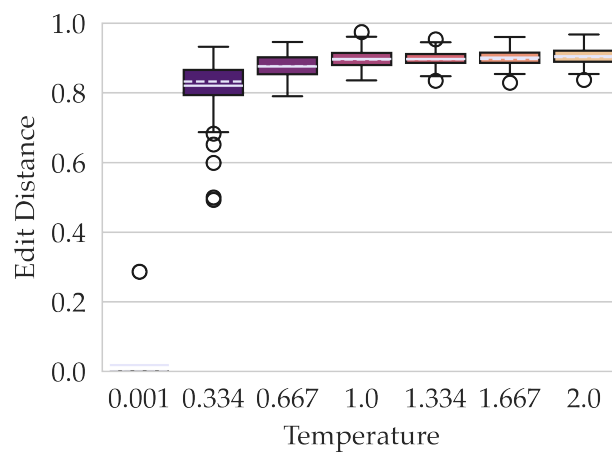
In general, the distribution plots (Figure 4.3) suggest that the LLM, for this specific prompt and model, does not produce more diversity at higher temperatures. Higher temperatures increase the odds of generating that diversity, but the measures indicate that it is not essential. In Figure 4.3b, the normalised edit distance measure shows that even at lower temperatures ($.334 < \tau < 1.0$), there is an immediate effect on diversity relative to the exemplar story. Although this does not tell anything about the meaning of the stories, it suggests that higher temperatures ($\tau > 1.0$) are not needed to achieve more diversity. Cosine similarity shows a weak negative trend, but also shows that there is great overlap in the range of outputs for each temperature value (Figure 4.3a). In other words, outputs at higher temperatures on the semantic level do not extend much further away from the exemplar story than outputs at lower temperatures.

In theory, given enough time and space, we should be able to generate outputs from every region of the embedding that the context of the prompt and model can be generated, even at lower temperatures. However, when drawing limited samples from the LLM, as is standard in any real-world scenario, we do observe some exploration at higher temperatures, as is shown in Figure 4.4. At the very least, it appears that temperature increases the chance of finding novelty more quickly.

Another interesting observation is that it seems, relative to the exemplar story, that the generation process moves in a particular direction (Figure 4.4). This suggests some merit to the link we draw with prototype and exemplar theory and our approach to evaluation, as it shows how LLM outputs change relative to the exemplar.



(a) Cosine similarity of stories with the exemplar per temperature value.



(b) Normalised edit distance of stories with the exemplar per temperature value.

Figure 4.3: Both (a) and (b) suggest that higher temperatures do not imply more diversity on the semantic or lexical level, or that they further extend the range of possible outputs in the current context.

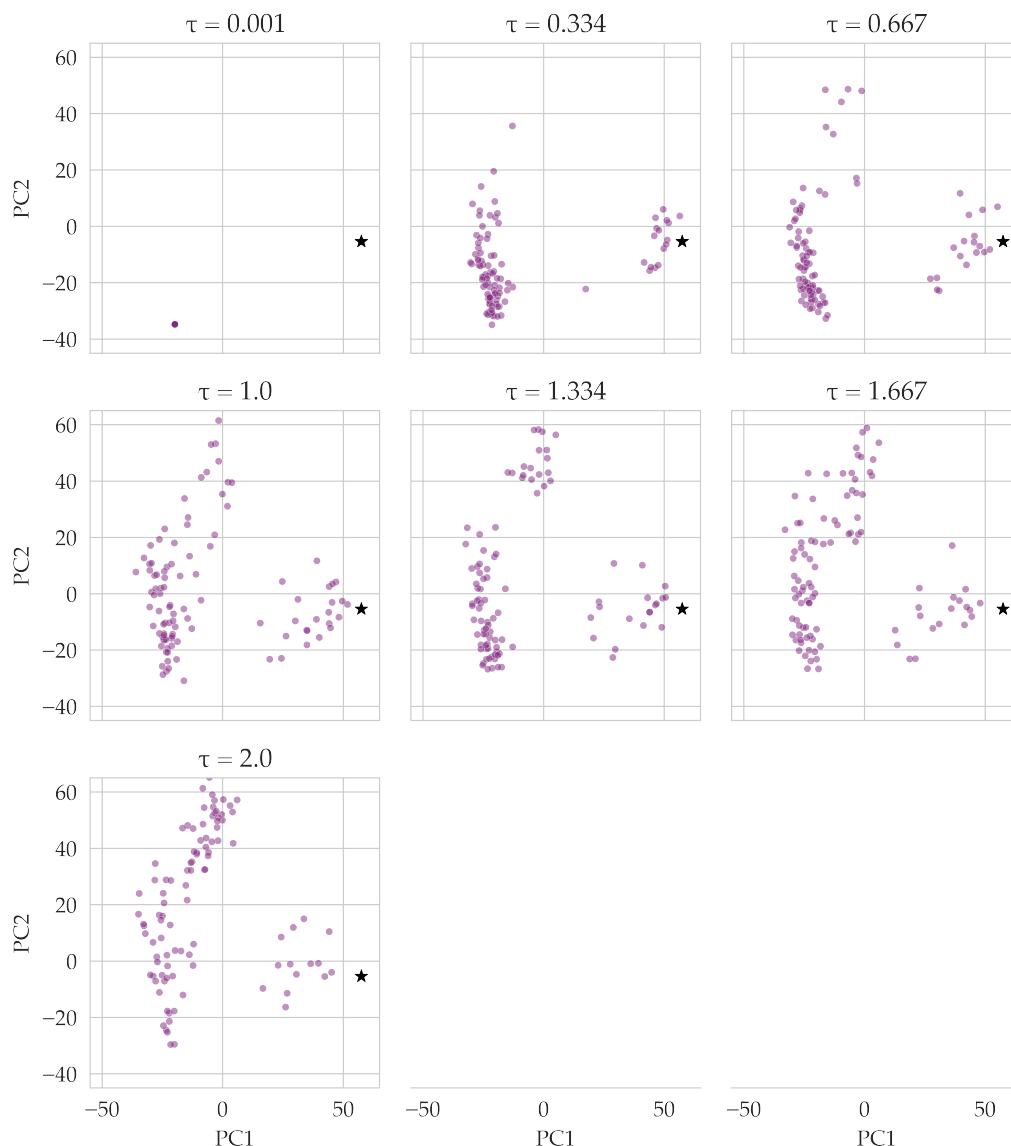


Figure 4.4: Here, we show the PCA projections of 100 stories per temperature generated using Llama 2 Chat 70B and Prompt 4.3.1. While we visually observe that an increasing temperature seems to explore a larger region of the embedding space with a small number of samples, it does not imply that we access a broader slice of the model’s output space. We merely increase the chance of generating more diversity. ★ denotes the exemplar story.

4.4.2 Experimental Analysis with Human Evaluation

The computational analysis is limited because it does not inform about the meaning of the stories. It might be that even though our measures show clear limitations in terms of diversity and overlap in output distributions, stories generated at higher temperatures are indeed creative in ways not captured by the computational analysis. To further investigate this, we designed a small creativity evaluation experiment with non-expert human participants. In the experiment, participants evaluate stories according to the four necessary conditions mentioned in the Background section. Specifically, we ask them to compare each story for novelty and typicality against the exemplar and judge the story for its individual cohesion and coherence. To generate the stories, we again use Llama 2 Chat 70B with the same prompt (Prompt 4.3.1), with the same setup described in the Exemplar Story section. For each of the 7 temperature values $\tau \in [.001, .334, .667, 1.0, 1.334, 1.667, 2.0]$, we randomly generate 5 unique stories that are at least 300 tokens long. At the lowest temperature setting $\tau = .001$, the model could generate only one other unique story besides the exemplar story; as such, we have a total of 31 stories for the creativity evaluation. Figure 4.5 shows the exemplar story, and Figure 4.6 and Figure 4.7 show two of the generated stories, at $\tau = 1.0$ and $\tau = 2.0$ respectively.

Experimental Setup

For this study, we recruited 36 participants online and from the Computational Creativity module at the University of Kent. Due to the ethics review, we did not collect any personal demographic data, but all participants were 18 years or older, studying in the UK for an undergraduate or master's degree in Computer Science. Each participant was asked to evaluate 5 LLM-generated stories using a questionnaire. They were told upfront that the stories were generated by an LLM, but they were unaware of the specific model or that the stories were sampled from different temperature values. The participants were given the definitions of the four necessary conditions for creativity, followed by the exemplar story (Figure 4.5), and a two-story practice evaluation. The practice stories are the same for all participants and given in random order. After practice, they were again presented

Once upon a time, there was a young girl named Maria. She lived in a small village nestled in the rolling hills of the countryside. Maria was a curious and adventurous child, always eager to explore the world around her.

One day, while wandering through the village, Maria stumbled upon a hidden path she had never seen before. The path was overgrown with weeds and vines, and it looked as though it hadn't been used in years.

Without hesitation, Maria decided to follow the path. She pushed aside the branches and vines, and began to walk down the winding trail. The path led her through the forest, and the trees grew taller and closer together as she walked.

As she continued on, Maria began to notice strange markings on the trees. They were unlike anything she had ever seen before - strange symbols and patterns that seemed to glow in the fading light of day.

Maria's curiosity grew with each step, and she quickened her pace. She knew that she had to find out what these markings meant.

After a while, the path opened up into a clearing. In the center of the clearing stood an enormous tree, its trunk twisted and gnarled with age. The symbols and patterns that Maria had seen on the other trees were etched into the bark of this tree as well, but they were even more intricate and complex.

Maria approached the tree, her heart pounding with excitement. As she reached out to touch the bark, a voice boomed from within the tree.

"Who dares to disturb the resting place of the ancient ones?" it said.

Maria was startled, but she stood her ground. "I'm just a curious child," she said. "I didn't mean to disturb anyone."

There was a long silence, and then the voice spoke again. "You have been chosen to receive a great gift," it said. "Are you willing to accept it?"

Maria nodded eagerly, and the tree began to glow with a soft, golden light. The symbols and patterns on the bark began to move and shift, and Maria felt a strange energy coursing through her body.

When the light faded, Maria found that she had been given the power to communicate with the spirits of the forest. From that day on, she used her gift to help the villagers and the forest creatures, and she lived a long and happy life filled with wonder and adventure.

The end.

Figure 4.5: Exemplar Story ($\tau = 0.0$)

The sun was setting, and a group of friends had gathered on the beach to watch it go down. They sat in comfortable silence, each lost in their own thoughts, as the sky transformed into a canvas of pink, orange, and purple hues.

Suddenly, a young boy appeared, walking towards them from the water's edge. He was dripping wet, and his eyes were wide with excitement.

"What did you find, little man?" one of the friends asked, as the boy approached.

"I found a message in a bottle!" the boy exclaimed, holding up a glass bottle with a piece of paper inside.

"To whoever finds this,

I hope this message finds you well. My name is Sarah, and I'm a pirate. I've been sailing the seas for years, searching for treasure and adventure. I've seen so many amazing things, from dolphins playing in the waves to fierce storms that threaten to sink my ship.

But the best part of my journey has been the people I've met along the way. I've made friends with merchants, fishermen, and even other pirates. We've shared stories, laughed together, and sometimes, we've even saved each other's lives.

I'm writing this message in a bottle, hoping that it will find its way to someone who will appreciate the adventure and camaraderie that comes with life on the high seas.

If you're reading this, I hope you'll take a moment to think about the amazing things you've experienced in your own life. And if you're feeling adventurous, maybe you'll even decide to set sail and see what wonders the world has in store for you.

Fair winds, Sarah the Pirate"

The group of friends sat in silence for a moment, taking in the message. Then, they all started talking at once, sharing their own stories of adventure and camaraderie. They laughed and reminisced, feeling grateful for the chance to share their experiences with each other.

As the sun dipped below the horizon, the group of friends raised their glasses (filled with whatever drink they had on hand) in a toast to Sarah the Pirate, to adventure, and to the magic of messages in bottles. And who knows, maybe one day, they'll even set sail themselves, following in Sarah's footsteps and creating their own stories of adventure and friendship.

Figure 4.6: Example Story ($\tau = 1.0$)

Once upon a time, there was a beautiful young woman named Anastasia. She was smart, kind, and full of life, with long brown hair, deep blue eyes, and a bright, radiant smile. Ana was a gentle soul, loved by all who knew her, and she spent her days helping those in need.

One day, while out on a walk, Ana stumbled upon a hidden path she had never seen before. Curious, she decided to follow it, and it led her to a clearing surrounded by tall trees. In the center of the clearing was a beautiful, ancient-looking statue of a woman. Ana felt drawn to the statue and walked closer to examine it. As she reached out to touch it, she suddenly felt a strange sensation, like a jolt of electricity running through her body.

When the sensation passed, Ana found that she was no longer alone. A handsome young man stood before her, dressed in clothing unlike anything she had ever seen. He introduced himself as Mikhail, and Ana was struck by his kind eyes, sharp jawline, and charming smile. Mikhail told Ana that he was a guardian, sent to watch over the statue and protect it from those who would misuse its power.

Ana was fascinated, and asked Mikhail to tell her more. He explained that the statue was a symbol of love and hope, and that it had been placed there by a wise queen centuries ago. The queen had imbued the statue with magical powers, which would grant the deepest desires of those who approached it with a pure heart. But, Mikhail warned, the statue also had the power to bring great destruction if used by those with malicious intent.

Ana was amazed by the story, and felt honored to have been chosen to meet Mikhail and learn about the statue's secrets. She knew that she had always had a pure heart, and she longed to help others, so she asked Mikhail if she could stay and help him protect the statue.

Mikhail agreed, and together they worked to keep the statue safe from those who would misuse its power. Over time, Ana and Mikhail grew closer, and they realized they had fallen in love.

Years passed, and their love only grew stronger. They were happy, but they knew their time together was limited, for Mikhail's duty as guardian was eternal, and Ana could not stay with him forever.

One day, a wicked sorcerer appeared, seeking to use the statue's power to gain control over the kingdom. Mikhail fought bravely, but the sorcerer was too strong. Ana knew she had to do something, and with a burst of courage, she stepped forward and touched the statue.

The sorcerer was repelled, and as he fled, Ana realized that she had been granted the statue's power. She felt a surge of strength, and knew that she could now help Mikhail protect the statue, and the kingdom, from any harm that might come.

Ana and Mikhail spent their days watching over the statue, using its power for good, and spreading love and hope throughout the land. They knew they would always be together, bound by their love for each other, and their duty to protect the magical statue and all it represented.

And so, they lived happily ever after, surrounded by the magic and beauty of the ancient statue, and the love that they shared.

Figure 4.7: Example Story ($\tau = 2.0$)

with the exemplar story before starting the actual creativity evaluation. The exemplar story and definitions were accessible throughout the survey. The study was approved by the Central Research Ethics Advisory Group of the University of Kent.

Each participant was randomly assigned 5 of the 31 stories. The evaluation consists of 4 five-point scale questions; two about novelty and typicality *compared against the exemplar story*, and two about the stories' text cohesion and coherence. The exact definition for each conditions is as follows:

- **Novelty:** A story is novel if it has something new or unusual.
- **Typicality:** A story is considered typical if it exhibits essential characteristics of a story.
- **Cohesion:** A story is cohesive if the use of words and sentences are grammatically consistent. Characters, events, and actions are used appropriately.
- **Coherence:** A story is coherent if it is easy to follow and understand its meaning.

The questions about novelty, typicality, and cohesion are on a five-point scale from (1) not at all _____, to (5) completely _____. Notably, coherence is on a five-point scale from (1) very easy to (5) very difficult to understand and follow. As a result, lower scores indicate high coherence, and vice versa. The survey distribution procedure was set up such that each story received at least five evaluations by different participants.

Concluding the survey, we posed 5 post-evaluation questions about their experience with reading, writing, and reviewing stories, their stance towards Generative AI and its use for creative tasks, and how they approached the evaluations.

Practical Considerations

There are several practical considerations regarding the experimental design that limited the study in various ways. The key limitation of this study is its size; we recruited 36

participants, who could each only evaluate a limited number of stories. We selected 5 unique stories per temperature set generated for the computational, relying on the distribution of the LLM to provide an appropriate sample. Ideally, you would generate many stories and then take a representative sample, but we do not know what number is sufficient for our purpose. Similarly, we only considered one prompt and its exemplar. Ultimately, we decided on, in our opinion, the most minimal and neutral instruction to perform the task (Prompt 4.3.1).

In this kind of setup, there is a risk of practice and order effects. To mitigate these issues, we asked each participant to first do a practice evaluation, making sure that the participant is fully aware of the task and what is expected of them. As mentioned, we randomly assigned stories to the participants, and did not observe any ordering effects in the analysis. We considered using pairwise comparison ranking, and presenting stories in pairs and evaluating those relative to each other, to further mitigate these effects. While this method is effective towards that purpose, it does not scale very well; even with just 31 stories, it would require a large number of participants for robust and meaningful outcomes.

Another limitation of this study is that cohesion and coherence are, in fact, important factors for novelty and typicality (Ritchie, 2007; Ventura, 2016). These LLMs produce fluent and grammatically correct language; this does not appear to be an issue. The stories generated by Llama 2 Chat 70B are simple stories featuring only basic elements, and are not literary masterpieces. We are interested in a relative evaluation to see how temperature changes creativity as it increases only for a simple task; thus, we opted to use non-experts who are not preconditioned to higher expectations. For that same reason, we designed the survey around specific criteria and did not ask participants directly about the creativity of the stories.

Analysis

We use linear mixed-effects models (Seabold & Perktold, 2010) to control for the random effect of assigning stories to different participants, and Cronbach's α to test inter-rater

Table 4.1: Evaluation Ratings and Reliability

	Mean	Std. Dev.	Cronbach's α
Novelty	3.12	1.17	.547
Typicality	3.20	.987	.664
Cohesion	3.69	.959	.596
Coherence [†]	2.13	1.17	.749
Mean	—		.639

† lower indicates higher coherence; higher indicates incoherence.

reliability. The approach is as follows: First, we determine if there is a significant effect using Satterthwaite's method (Satterthwaite, 1941). Secondly, the model's estimate $\hat{\beta}$ (slope) determines the direction of the effect, i.e. positive or negative, and the rate of change. Finally, we test how much variance is explained by the model to determine the magnitude using conditional R^2 (Nakagawa & Schielzeth, 2013), denoted by R_c^2 , which considers both fixed and random effects. The more variance is captured by the linear mixed-effects model, the stronger the correlation. Additionally, we report the marginal R_m^2 , which is the variance explained by the fixed effects.

Quantitative Results

The primary results of the analysis focus on the effect of temperature on the creativity of stories. Regarding inter-rater reliability, on average, we find $\alpha = .639$ (Table 4.1), which is just short of the acceptable level of .7, but still indicates some degree of agreement, especially so because we are dealing with non-expert judges who, in general, have lower inter-rater reliability scores (Lamb, Brown & Clarke, 2016). Regarding temperature, we observe two effects (Table 4.2 and Table 4.3): a weak positive correlation between novelty and temperature ($\hat{\beta} = .308$, $SE = .138$, $R_c^2 = .385$), and a moderate positive correlation between incoherence and temperature ($\hat{\beta} = .240$, $SE = .122$, $R_c^2 = .646$). This implies a

trade-off between novelty and coherence. While higher temperatures seem to positively affect the chance of finding novel stories, they become less coherent. Although coherence is rated lower, the mean is 2.13 (where 1 is very coherent and 5 is very incoherent), meaning that the stories, overall, were easy to understand. The negative correlation of coherence with temperature is as expected (Holtzman et al., 2020). Both effects have relatively shallow slopes, indicating that the rate at which novelty and coherence change, as temperatures get higher, is low, indicating only a small effect.

The secondary results relate the creativity evaluation to the computational metrics in the Computational Analysis section. Note that perplexity is a metric relative to the model, and cosine similarity and normalised edit distance are both relative to the exemplar story. Here, we observe three effects for perplexity, cosine similarity and normalised edit distance in Table 4.2 and Table 4.3. First, we see that perplexity has a weak to moderate positive correlation with novelty ($\hat{\beta} = .730$, $SE = .273$, $R_c^2 = .370$). This is interesting because it suggests that perplexity, although a measure for model quality, could also indicate novelty, at least to some extent. Yet, it is also unsurprising because, indirectly, higher perplexity is on average a result of higher temperature. There is likely a similar cause for the weak to moderate positive effect between normalised edit distance and novelty ($\hat{\beta} = 2.01$, $SE = .731$, $R_c^2 = .378$). Finally, there is a weak to moderate positive effect between cosine similarity and cohesion ($\hat{\beta} = 2.09$, $SE = .716$, $R_c^2 = .431$), which is not unexpected as higher temperatures produce outputs with greater variety, subsequently, increasing the likelihood of producing non-cohesive texts. More interestingly, this suggests that stories closer to the exemplar are more cohesive. In turn, this implies that when the semantic distance between a story and the exemplar is large, more features related to cohesion are lost. The slopes of the analysis with the computational metrics are steeper, especially the latter two, which have a high rate of change, looking at Figure 4.3a and Figure 4.3b. This could be due to the large difference between the lowest and highest temperatures.

Table 4.2: Statistical Analysis of the Creativity Evaluation and Computational Metrics I

	Temperature			Perplexity		
	$\hat{\beta} \pm \text{SE}$	R_c^2	R_m^2	$\hat{\beta} \pm \text{SE}$	R_c^2	R_m^2
Novelty	.308 $\pm$.138*	.385	.152	.730 $\pm$.273**	.370	.158
Typicality	-.095 $\pm$.118	.339	.125	-.205 $\pm$.237	.345	.126
Cohesion	-.181 $\pm$.112	.406	.161	-.183 $\pm$.226	.392	.151
Coherence [†]	.240 $\pm$.122*	.646	.283	.194 $\pm$.251	.627	.274

† positive correlation denotes a negative effect, i.e. higher temperatures are less coherent

* $p < .05$ ** $p < .01$

Table 4.3: Statistical Analysis of the Creativity Evaluation and Computational Metrics II

	Cosine Similarity			Normalised Edit Distance		
	$\hat{\beta} \pm \text{SE}$	R_c^2	R_m^2	$\hat{\beta} \pm \text{SE}$	R_c^2	R_m^2
Novelty	-1.08 $\pm$.913	.338	.129	2.01 $\pm$.731**	.378	.161
Typicality	-.987 $\pm$.771	.336	.128	-.662 $\pm$.632	.336	.126
Cohesion	2.09 $\pm$.716**	.431	.187	-.663 $\pm$.608	.384	.152
Coherence [†]	1.03 $\pm$.788	.628	.278	.750 $\pm$.664	.629	.276

† positive correlation denotes a negative effect, i.e. higher temperatures are less coherent

* $p < .05$ ** $p < .01$

Qualitative Results

In the post-evaluation questions, the participants reported on average some familiarity with generative AI, and were moderately comfortable with AI being used for creative work. Most participants reported very little to no experience with writing and reviewing stories. Those who reported having done writing were mostly relying on educational experience. We specifically asked the participants about their approach for evaluating

stories to gain further insights into what specifically they focused on. Participants reported different ways in which they approached the evaluation. Some focused on structure, others aimed to grab the general idea, and combined and compared the endings. Occasionally, the participant chose one specific factor, such as that most stories were adventure-based, and took that as the typical trait. Another example is that when a story changes genre halfway, for example, to fantasy, it is considered more novel. In line with this, participants did report that it was quite difficult to separate details in the stories as they went through the evaluation, and that it was quite difficult to just use the four conditions as there were different characteristics in the story they had to weigh. One participant reported a lack of cohesion, and that sometimes it felt that sentences were skipped, misplaced, or there was something missing. These reports suggest that for each criterion, different things are expected and valued in LLM-generated stories, and that overall poor quality may have made the evaluation a challenge.

4.5 Discussion

Overall, we do not find compelling support for the “temperature as the creativity parameter” claim across the conditions for creativity in this investigation. While we see some positive effects for evaluating novelty, the rest of the results do not show any significant findings, meaning that there is not enough evidence that temperature as the creativity parameter enables creativity in LLMs.

The key to the evaluation methodology in this chapter is the idea of the exemplar, inspired by Prototype and Exemplar theory, as a point of departure for evaluation. As mentioned, we observe that the exemplar sits on the edge of the PCA projection (Figure 4.4). While this suggests that the exemplar is a reasonable baseline, it is necessary to further examine the merits of the link we draw between prototype and exemplar as an evaluation methodology. Further investigation is needed to see if this phenomenon generalises and emerges when applying this methodology to different domains, prompts, and models.

In the qualitative analysis, the participants reported using different internal criteria to evaluate novelty and typicality and found it difficult to separate details between stories during the evaluation. While this is nothing new, these reports complicate the robustness of our findings. Repeating the experiment with experts, better stories, and creative writing rubrics is important for future work. Moreover, an expert’s depth of understanding of the domain could identify other, more advanced desiderata. On the other hand, the stories generated with this simple prompt are very basic, and experts might judge them all as poorly written. This might complicate evaluation; as such, using non-experts might still be adequate and more readily achievable. Designing a robust, large-scale experiment to investigate the effect of temperature on creativity, given the vast amount of variation that LLMs can produce, is difficult and needs careful consideration.

Another limitation is the visual inspection of the PCA to see if the temperature changes access to different regions of the output space. In future work, a more robust approach would be to increase the sample size or apply clustering techniques, providing stronger justification for concluding that the temperature parameter has a limited effect on the range of generated artefacts.

4.6 Conclusion

In summary, this chapter presents a two-fold empirical study to investigate how temperature affects the creativity of LLM-generated stories. The novel methodology proposed in this chapter leverages the idea of the exemplar, inspired by prototype (Rosch, 1973; Rosch et al., 1976) and exemplar theory (Medin, Altom & Murphy, 1984), as a point of departure for evaluation, mitigating variance introduced by the prompt. The findings indicate that:

- When generating a limited number of samples, temperature increases the generated variety, but overall, it does not enable access to a larger semantic space.
- We only observe a weak positive correlation between novelty and temperature, and

a moderately negative correlation between coherence and temperature, suggesting a novelty and coherence trade-off mediated by temperature.

To conclude, the influence of temperature is far more nuanced and weak than “the creativity parameter” claim suggests. This study shows that it cannot be considered a universal creativity parameter. The trouble with using temperature as a means to generate a more diverse set of outcomes and thus control the creativity of the LLM comes at the cost of coherence, as shown in Table 4.2. An important and useful research direction is to explore how good LLMs are at generating such a diverse set, and more importantly, how we can increase diversity while maintaining quality. The next chapter investigates the loss of diversity due to instruction-tuning, as was first shown by (Hämäläinen, Tavast & Kunnari, 2023), specifically for a narrative generation task. Currently, we need to rely too heavily on chance to produce relevant and creative output. These recommendations should steer us towards more informed and useful creative behaviours of LLMs.

Chapter 5

Conformative Decoding: Minding the Diversity Gap

This chapter is based on: Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2025). Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models [Preprint, Under review]. <https://arxiv.org/abs/2507.20956>

5.1 Introduction

Chapter 4 showed that raising the temperature parameter allows some exploration of the embedding space. A natural follow-up is to ask how diverse the outputs of large language models (LLMs) for creative tasks are, and if there is room to improve. This is especially true for instruction-tuned LLMs, which are known to generally produce less diverse outputs than their base models (Hämäläinen, Tavast & Kunnari, 2023; Chu, Chen & Nakayama, 2024; Kirk et al., 2024; Le Bronnec et al., 2024; Shypula et al., 2025). Instruct models may have a bias towards certain answers or topics: for example, prompting an

LLM to write about an “evil empire” may disproportionately generate outputs related to Star Wars. However, following instructions and engaging in conversation is an extremely useful application of LLM. The base model can only rely on pre-existing token patterns in the training data to generate an output of the desired artefact class. Few-shot prompting (Brown et al., 2020) can push the base model to more consistent behaviour but requires a specific (diverse) set of examples. Such a set of the desired artefact class may not be available, or the user may not know what they are looking for. Even with examples, it is not trivial to understand how the LLM uses those examples and, subsequently, steers the generation process.

Output diversity is a desired characteristic of LLMs and other generative models (Tevet & Berant, 2021). This is not limited to mundane interactions, such as “Any tips for things to do in Edinburgh?” but also applies to more complex and creative tasks, such as narrative generation. It is valuable for an LLM to respond in different ways, not only to avoid repetitiveness in conversations but also to enable (creative) exploration. Diversity (as *variety, divergence, and experimentation*) is a key component of creativity identified by Jordanous and Keller (2016) in their analysis of relevant academic literature spanning over sixty years. Having a diverse set of options enables consideration of multiple perspectives before selecting the most suitable output (Grace & Maher, 2016; Ibarrola & Grace, 2024). Experiments also suggest that humans collaboratively writing with LLMs leads to reduced diversity in the produced content (Padmakumar & He, 2024). The loss of diversity due to instruction-tuning might have implications in applications such as agent-based simulations (e.g., Park et al., 2023), employing the LLM-as-a-Judge (Zheng et al., 2023) for evaluation or generating participant responses, or for inducing personalities (Jiang et al., 2023b), but also for generating synthetic instruction-tuning datasets for training other LLMs.

In this chapter, I investigate whether instruction-tuning reduces the output diversity of various open-source and open-weight LLMs when applied to an open-ended narrative generation task. Specifically, I examine OLMo 0724 7B (Groeneveld et al., 2024) and OLMo 2 7B 1124 (Team OLMo, 2025) in more detail to determine if the “diversity gap” can be attributed to different stages of preference tuning. Through narrative generation

tasks, I demonstrate the existence of a diversity gap due to instruction-tuning in five LLMs. The results show that Direct Preference Optimisation (DPO; Rafailov et al., 2023) is primarily responsible for the diversity gap, as it considerably reduces the output diversity across the generated narratives. Motivated by these findings, I propose conformative decoding, a strategy designed to reintroduce diversity by attenuating the instruct model using the distribution of its more diverse base model, thus allowing the usefulness of instruction-tuned LLMs while enabling more diverse outputs. The results show significant improvements while maintaining quality, except for Llama 3.1 8B, which is an interesting deviation from this trend, which I discuss in more detail in section 5.5.

5.2 Background

Diversity in generative AI outputs is receiving more attention, especially in evaluation (e.g., Pillutla et al., 2023; Le Bronnec et al., 2024). For LLMs, various methods and decoding strategies exist to make them into more diverse generators (e.g., Vijayakumar et al., 2018; Holtzman et al., 2020; Li et al., 2023). These recent decoding methods can produce text that is of high quality or at least as diverse as human text. However, one question that has not been addressed is whether such methods can overcome the loss of diversity due to instruction-tuning. In this section, I discuss what diversity is, how to measure it, and review the efforts to make LLMs more diverse.

5.2.1 What is diversity?

Diversity is a complex and dynamic concept. It has different meanings in different research areas, shaped by the goals and challenges. Generally, diversity refers to the presence of differences or variety in some context along some dimension. More strictly, diversity is a quality observed in a set of objects; a single entity cannot be considered diverse in itself.

One view of diversity based on entropy is used in ecology to quantify biodiversity

(Leinster, 2021). A population is diverse because it has a wide range of varieties of life. Language is typically concerned with two kinds of diversity: lexical diversity and semantic diversity (Tevet & Berant, 2021; Stasaski & Hearst, 2022; Guo et al., 2024). The former is about style or form, while the latter is about content. Diversity plays an important role in creative domains; for example, in music, if a band is known for its diverse music, it usually means their songs are not confined to a single genre. Moreover, diversity is well-studied in information retrieval, specifically for recommender systems, which need diversity to satisfy users. However, users do not always desire the same kind of diversity; some music consumers may want diversity within genre, and others may want to explore beyond familiar genres (Robinson, Brown & Schedl, 2020). Moreover, diversity can be connected to unexpectedness, but in relation to user's expectations, rather than the diversity to the set of recommendations itself (Adamopoulos & Tuzhilin, 2014). This implies that diversity within a set does not guarantee finding a surprising and novel object.

When is a generative AI model diverse? Typically, when a model produces a broad and diverse set of appropriate outputs, it is more useful in different kinds of situations. A model that only generates images or texts in the same style or recommends the same songs repeatedly is not very diverse or useful. Another view is that diversity could be about representational diversity. This line of research investigates how generative models depict minority groups (Shihadeh & Ackerman, 2023a; Ackerman & Brown, 2024) or associate certain human characteristics with a certain demographic (Fraser, Kiritchenko & Nejadgholi, 2023; Shihadeh & Ackerman, 2023b). While I highly appreciate the valuable work being done on representation and diversity, this chapter focuses on linguistic diversity.

5.2.2 Measuring Quality and Diversity

Quality and diversity are two key criteria for assessing the creativity of outputs (Jordanous & Keller, 2016; Ibarrola & Grace, 2024). Diversity appears closely related to novelty (Runco & Jaeger, 2012), but distinct in that a diverse generator can generate novel objects,

but those do not necessarily contribute to diversity. Object chosen from a conceptual space (Boden, 1992) that is well explored should have high diversity, and some novel objects could still be found. Still, a diverse generator does not necessarily know how to go beyond the scope of its training data.

Reference-based Metrics

Evaluating generative AI models generally focuses on the same two axes: quality and diversity. There is usually a trade-off between the two (Le Bronnec et al., 2024). The metrics generally require embedding the generated artefacts in a latent space. A common way to measure quality is to use the Fréchet distance using the latent space of another model. For example, the Fréchet Inception Distance (FID) for measuring image quality uses the latent space of the InceptionV3 model (Heusel et al., 2017). Another method designed for assessing quality and diversity in generative models, respectively, is Precision and Recall using distributions (Sajjadi et al., 2018). The Improved Precision and Recall (Kynkäänniemi et al., 2019) metric works by drawing a hypersphere at each data point with a radius up to its k -th nearest neighbour to estimate manifolds for the generated samples Φ_g and human reference Φ_r . The following function determines if a point ϕ falls within the manifold Φ : it returns 1, or if not it returns 0:

$$f(\phi, \Phi) = \begin{cases} 1, & \text{if } \|\phi - \phi'\|_2 \leq \|\phi' - NN_k(\phi', \Phi)\|_2 \text{ for at least one } \phi' \in \Phi \\ 0, & \text{otherwise.} \end{cases} \quad (5.1)$$

where NN_k returns the k -th nearest neighbour $\phi' \in \Phi$ of ϕ . This functions counts The Precision and Recall metric is then defined as follows:

$$\text{Precision}(\Phi_g, \Phi_r) = \frac{1}{|\Phi_g|} \sum_{\phi_g \in \Phi_g} f(\phi_g, \Phi_r) \quad (5.2)$$

$$\text{Recall}(\Phi_g, \Phi_r) = \frac{1}{|\Phi_r|} \sum_{\phi_r \in \Phi_r} f(\phi_r, \Phi_g) \quad (5.3)$$

Precision is the probability that a point in Φ_g falls within the support of Φ_r , and vice versa for Recall.

Another measure specifically designed for text is MAUVE (Pillutla et al., 2021), which also compares a generated and ground truth distribution, but by means of the KL divergence D_{KL} in a quantised embedding space. MAUVE has been found to align with human judgements, and its measurements are consistent regardless of text length, model size, and decoding strategy (Pillutla et al., 2023; Pimentel, Meister & Cotterell, 2023). Similar to Precision and Recall, MAUVE distinguishes two kinds of errors: Type I, where the generated text Q places mass on text that is unlikely under the human reference P (quality), and Type II, where Q cannot generate text that is plausible under P (diversity). These are quantified as $D_{\text{KL}}(Q \parallel P)$ and $D_{\text{KL}}(P \parallel Q)$, respectively. MAUVE is the area under the *divergence curve* (Equation 5.5) where a large value means that Q is closer to P , and when $P = Q$ the MAUVE score is 1.

$$\mathcal{F}(P, Q) = \left\{ \left(D_{\text{KL}}(Q \parallel R_\lambda), D_{\text{KL}}(P \parallel R_\lambda) \right) : \lambda \in (0, 1) \right\} \quad (5.4)$$

$$\begin{aligned} \text{MAUVE}(P, Q) = \text{AUC} \Big(\Big\{ (\exp(-cx), \exp(-cy)) : (x, y) \in \mathcal{F}(P, Q) \Big\} \\ \cup \{(1, 0), (0, 1)\} \Big) \end{aligned} \quad (5.5)$$

where c is a scaling constant and $R_\lambda = \lambda P + (1 - \lambda)Q$ is an approximation or softening, since in practice, the divergence curve without softening is infinite, unless $P = Q$ (Pillutla et al., 2021). Essentially, R_λ is a series of distributions interpolating P and Q to compute the divergence frontier $\mathcal{F}$. Pillutla et al. (2023, Figure 1) provide an illustration of the divergence frontier and the interpolation using R_λ . Hence, MAUVE is designed to capture both quality and diversity since it takes both Type I and Type II errors into account, but it can be difficult to disentangle the two. A low MAUVE score could be due to a high Type II error caused by a lack of variety (for example, repetitiveness, limited vocabulary), or a high Type I error caused by poor quality (in other words, badly constructed sentences with a wide vocabulary). MAUVE is recommended when both diversity and quality

are the objective (Pillutla et al., 2023). A higher MAUVE score means that the generated distribution Q is closer to the reference distribution P .

Automated approaches that focus specifically on measuring diversity have received less emphasis than those that focus on measuring quality. As mentioned, *Recall* is most commonly used to assess diversity. However, having a human reference that is itself considered diverse is essential so that if there is good coverage of the human reference by the generated samples, the generated set must also be diverse. This approach is taken by Hämäläinen, Tavast and Kunnari (2023) and Le Bronnec et al. (2024) to conclude that instruction-tuning lowers output diversity.

Reference-free Metrics

The metrics mentioned in the previous sections have two significant limitations. First, they all require a large sample size from the model to get reasonable results, and second, they require a large and diverse ground truth set. In reality, we often want to evaluate small sets of generated samples, and a diverse ground truth dataset is generally hard to obtain. Two measures specifically designed for diversity that do not have these drawbacks are the *Vendi Score* (Friedman & Dieng, 2023) and *Truncated Entropy* (Ibarrola, Lawton & Grace, 2024). Both of these metrics adopt an entropic view of diversity, but each takes a different approach.

The Vendi Score (Equation 5.6) of N objects is defined as the exponential entropy of the eigenvalues λ of similarity matrix $\mathbf{K}/N$, where $\mathbf{K} \in \mathbb{R}^{N \times N}$ is a positive semidefinite matrix, meaning all the eigenvalues are non-negative. The positive semidefinite similarity function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is required to have $k(x, x) = 1$ for any object x , and $k(x, y) = k(y, x)$ for any two objects x and y .

$$\text{VS}_k(\mathcal{X}) \triangleq \exp \left(- \sum_{i=1}^N \lambda_i \log \lambda_i \right) \quad (5.6)$$

The Vendi Score is bounded between 1 and N (Friedman & Dieng, 2023, Section 7.4),

and requires selection of an appropriate similarity function k . If the elements in $\mathcal{X}$ are all identical, the resulting similarity matrix $\mathbf{K}$ is all ones, meaning entropy is 0, and the Vendi Score returns its minimum score of 1. If all the elements in $\mathcal{X}$ are completely different, the similarity matrix is the identity matrix, it returns the maximum score N . Hence, the intuition is that higher scores mean high diversity.

Truncated Entropy (TE) was originally introduced as Truncated Inception Entropy, as it relies on InceptionV3 to measure image diversity. Like FID, truncated entropy is based on the differential entropy of a multivariate normal distribution; hence, like FID, it assumes the samples in $\mathcal{X}$ are normally distributed in the latent space. As Ibarrola, Lawton and Grace (2024) note that the empirical approximation $\hat{\Sigma}$ of covariance matrix Σ is singular, and its determinant zero, when the number of samples N is less than the dimensions of the latent space, making it infeasible to compute the entropy for small sets of artefacts. They propose a truncated version of differential entropy (Equation 5.7) and consider only the set of N largest eigenvalues of $\hat{\Sigma}$ to make computation feasible. Here, the superscript index l denotes the latent space in which the artefacts are embedded. Originally designed to measure image diversity, its extension to CLIP included an experiment into text diversity (Ibarrola & Grace, 2024).

$$\text{TE}_l(\mathcal{X}) \triangleq \frac{N}{2} \log(2\pi e) + \frac{1}{2} \sum_{i=1}^N \log \lambda_i^{(l)} \quad (5.7)$$

The intuition for Truncated Entropy is similar to that for the Vendi Score, but unlike the Vendi Score, it is unbounded and can be negative when the eigenvalues are small.¹ Truncated Entropy does not use an explicit similarity function but relies on the latent space of an embedding model. However, diversity is measured by the distance between samples, which operates as an implicit similarity function. In that sense, the similarity is provided by the embedding model under the assumption of a normal distribution. This means that if Truncated Entropy is applied to an $\mathcal{X}$ with all identical elements, the distribution is highly concentrated and results in a low diversity score. When $\mathcal{X}$ consists

¹This happens when Truncated Entropy is applied to normalised vectors; these should be avoided.

of all dissimilar elements and is more widely distributed, the diversity score increases.

Arguably, the Vendi Score has some advantage over Truncated Entropy, as it allows different similarity kernel functions k to assess the desired flavour of diversity. Moreover, because the Vendi Score is bounded between 1 and N and does not rely on assumptions about the data, it is easier and more natural to interpret. However, Truncated Entropy can achieve comparable flexibility by varying the embedding space l .

Embedding-free Metrics

These metrics all require some kind of embedding in a latent space. Other metrics that have been used to measure diversity and quality and which do not require embeddings are n -gram diversity (Li et al., 2016), and for the latter, perplexity and Zipf's Law (Zipf, 1949). However, perplexity is not a good measure for text quality overall (Nadeem et al., 2020; Pillutla et al., 2021), instead it is better for measuring model quality given training data. The Zipf coefficient is more a measure of closeness to natural language than quality per se. While n -gram diversity is a standard diversity measure, it has some drawbacks. If the sequences are of different lengths, the diversity of a set can be overestimated if it contains longer sequences with many unique words and conversely underestimated for short sequences with many repeating occurrences (Friedman & Dieng, 2023, Section 9.4).

In this section, I have summarised metrics that compute the diversity and quality of the output of generative AI models. To summarise, the reference-based metrics require a ground truth to compare against the generated data, and rely on the assumption that the assessed properties are present in the human reference data. For diversity, there are two reference-free metrics: one uses a similarity kernel (Vendi Score) and another (Truncated Entropy) that uses the covariance matrix of the generated data to compute a diversity score. In the next section, I will discuss several approaches aimed at improving the output diversity of LLMs.

5.2.3 Improving the Output Diversity of LLMs

This section presents and expands the background discussed in subsection 2.2.2. Efforts to improve the output diversity of LLMs typically focus on decoding strategies. Decoding is the process of sampling a token from the next-token probability distribution given by the LLM on input. These strategies are generally divided into two flavours: sampling and search-based methods.

The easiest method to increase output diversity is temperature sampling, where increasing the temperature hyperparameter flattens the sampling distribution, but at the cost of quality (See chapter 4). Often, increasing temperature needs to be combined with a truncation strategy such as top- k (Fan, Lewis & Dauphin, 2018) or nucleus sampling (Holtzman et al., 2020) to prevent, or at least reduce, the risk of producing degenerate outputs. The latter strategy, nucleus sampling, is interesting as it produces good quality text that is on par with diversity in human texts (Holtzman et al., 2020). Other sampling strategies perform similarly, as measured using n -gram diversity, with temperature sampling scoring best (Meister et al., 2023). Sampling decoding strategies generally do not focus specifically on optimising output diversity.

Search-based methods use beam search (often combined with a sampling strategy) to track multiple candidates and score them based on probability over multiple steps. The goal is to find good sequences with a high but “hidden” probability that might be overlooked when decoding using token-by-token sampling only. An effective search-based method that yields more diverse sequences is Diverse Beam Search (DBS) (Vijayakumar et al., 2018). DBS considers a fixed number of groups of beams, each group lagging one time step behind the previous group. The objective function combines the joint probability of beams and the dissimilarity between groups of beams at each step. This results in more diverse outputs compared to standard beam search. Contrastive Decoding (Li et al., 2023) is another search-based strategy that uses an objective where a larger model (the expert) is penalised by a smaller model (the amateur) to prevent the expert from following bad language patterns that the amateur has learned. Li et al. (2023) show that

Contrastive Decoding leads to more diverse outputs compared to other search-based methods, but is less effective than nucleus sampling. Search can increase output diversity but also has a higher cost compared to sampling.

Another approach is to optimise the diversity of the output, informed by multiple models using an ensemble of LLMs (Tekin et al., 2024). Depending on the input, this method selects a sub-ensemble out of a large ensemble of LLMs that is likely to yield a diverse output. Subsequently, the probability distributions of the sub-ensemble are combined, which requires an ensemble learner (another neural network), to yield diverse outputs. Finally, there are prompt-based approaches to increase diversity in various domains (Chu, Chen & Nakayama, 2024; Chen et al., 2025). Still, these are model- (and task-) specific, while the focus in this chapter is on a model-agnostic approach.

5.3 The Diversity Gap

In this section, I investigate the existence of a “diversity gap” for various small LLMs performing a narrative generation task. In particular, I explore the extent of the diversity gap using various diversity metrics, the impact of different fine-tuning steps on diversity, and the impact of conversation templates.

5.3.1 Experimental Setup

Narrative Generation Task and Data

The task used for evaluating LLMs on diversity and quality in this chapter is based on the concept of a writing prompt, a short statement or question designed to spark and inspire a writer’s creativity. Such tools to promote creativity have been around for a long time in various forms, such as *Exquisite Corpse* or the *Oblique Strategies* (Eno & Schmidt, 1975). The r/WritingPrompts subreddit is designed as a space to foster creative writing

and productivity. People post a small premise of a story they would like to see written, and others supply stories as a response in the comments.

However, since LLMs entered the scene, it has become increasingly challenging to find sources of data untouched by LLM-generated data. This is potentially problematic since many quality and diversity metrics rely on a human-made ground truth. Fortunately, the Writing Prompts dataset (Fan, Lewis & Dauphin, 2018) comprises prompts and stories from the r/WritingPrompts subreddit² before the widespread adoption of LLMs. To assess quality and diversity, there needs to be sufficient human stories for each writing prompt; hence, I use a subset of the Writing Prompts dataset, selecting all instances from the training dataset that have at least 50 human responses. We manually filtered out prompts that were considered inappropriate or offensive, and responses that began with commentary of some sort (e.g., “This is my first post” or “Feedback welcome”). The subset consists of the first 50 human responses for each writing prompt, resulting in a total of 53 prompts and 2650 human responses. Since the Writing Prompts dataset is considered diverse (Fan, Lewis & Dauphin, 2018), the subset is a suitable ground truth for the diversity metrics when required.

Prompting and Incipits

Base and instruct models require different prompts to perform a task; therefore, there is a text completion prompt and a conversation prompt (Prompt 5.3.1). The base models are always prompted such that their natural response is to complete a story, denoted as (*comp*) in the results. In the case of instruct models, I either use the completion prompt or a prompt using the conversation template, denoted as (*chat*) in the results, to see if there is a difference in diversity between a completion and a conversation prompt for instruct models. The conversation prompt is similar to the completion prompt (Prompt 5.3.1) and differs only in the instruction, replacing “The story is as follows:” with “Write a story.” and using a chat template.

In particular, consistent behaviour between the base and instruct models is essential. The

²<https://www.reddit.com/r/WritingPrompts/>

Writing prompt: [WP]

The story is as follows: [incipit]

Prompt 5.3.1: The exact prompt for text completion, [WP] and [incipit] are replaced with the writing prompt and the first 20 tokens of a human response, respectively. For the conversation prompt, “The story is as follows: ” is replaced with the instruction “Write a story” and apply the model-specific chat template, leaving the rest, including the supplied incipit, unchanged.

generated outputs must be of the correct artefact class: stories. Instruct models, using the conversation prompt, have no problem performing the task. Generating stories using the completion prompt is much harder, especially for base models, which often output degenerate, repetitive text or items that are not a story. Example stories for each model can be found in the appendix C.4.

To alleviate this unwanted behaviour, I use incipits of the human stories in the dataset to naturally guide the models into their completion task. These incipits introduce a bit of diversity but ensure the models properly perform the task on both prompts. The incipit size is 20 tokens. For consistency, the tiktoken library with the o200k_base encoding is used to truncate the human responses into incipits for all models. Few-shot prompting (Brown et al., 2020) is beyond the scope of this chapter, as it introduces too much natural variation, making consistent comparison difficult. The completion and conversation naturally have some differences, but I aim for the prompts to be as simple and similar as possible. See Prompt 5.3.2 and Prompt 5.3.3 as examples of writing prompts along with incipits.

Models

For practical reasons, the experiments focus on smaller open-weights and open-source models. In particular, I look at Gemma 2 9B (Gemma Team, 2024), Meta Llama 3.1 8B

```
<|endoftext|><|user|>
```

Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ...

Write a story.

```
<|assistant|>
```

Most people assumed that I most love night time after all how could a mage like me not love the

Prompt 5.3.2: An example prompt and its incipit (in bold) using the chat template of OLMo model family.

(Grattafiori et al., 2024), Mistral 7B v0.3 (Jiang et al., 2023a), OLMo 7B 0724 (Groeneveld et al., 2024), and OLMo 2 7B 1124 (Team OLMo, 2025). The OLMo families are fully open-source and provide intermediate checkpoints of each fine-tuning step, giving the opportunity to see how much diversity is lost in the Supervised Fine-tuning (SFT), Direct Preference Optimisation (DPO; Rafailov et al., 2023), and Reinforcement Learning with Verifiable Rewards (RLVR; Lambert et al., 2025) steps. As such, I am running additional experiments to investigate each fine-tuning step of the OLMo models in further detail. Every intermediate checkpoint is an instruct model; hence, each is evaluated with both the completion and the conversation prompt, as in the main experiment, while the base model is evaluated with the completion prompt only. All experiments use nucleus sampling ($p = .95$) with temperature $t = 1$. For each writing prompt in the Writing Prompts subset and each configuration: Base (*comp*), Instruct (*comp*) and Instruct (*chat*), I generate 50 samples using 50 incipits, up to a maximum of 500 tokens per story.

```
<|endoftext|><|user|>
```

Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ...

```
Write a story.
```

```
<|assistant|>
```

```
``This is very, very wrong.``
```

Erwan dumped the body in the cart. Turning

Prompt 5.3.3: An example prompt and its incipit (in bold) using the chat template of OLMo model family.

Evaluation Metrics

I evaluate the generated text using a range of quality and diversity metrics (lexical and semantic). To measure lexical diversity, I use Vendi Score (Equation 5.6) with n -gram counts for $n \in \{1, 2, 3, 4\}$ as features to compute the similarity matrix, denoted as $VS_{n\text{-gram}}$. For semantic diversity, I use the Vendi Score and Truncated Entropy (Equation 5.7) applied to text embeddings (see details below), denoted as VS_{jina} and TE_{jina} , respectively. Here, it is important to use raw embeddings for TE_{jina} as normalised features yield negative entropy, which is unintuitive to interpret. This happens when eigenvalues are very small, meaning the second term of Truncated Entropy dominates the first term.

I use Improved Precision and Recall (Kynkäänniemi et al., 2019) using $k = 3$ to compute Precision and Recall as quality and diversity metrics, respectively. High precision suggests realistic and accurate generated texts, while high Recall suggests good coverage of the ground-truth reference, implying high diversity. Furthermore, I evaluate the stories

using MAUVE with the default settings and follow best practices³. I use the text-matching embedding model mentioned below to compute the text features instead of the default GPT2-large. Note that I compute both MAUVE and Improved Precision and Recall across all prompts since these methods require many samples to give an accurate measurement, as opposed to the Vendi Score and Truncated Entropy, which are applied on a per-prompt basis.

To obtain the text embedding features, I use the `jinaai/jina-embeddings-v3` model (Sturua et al., 2024). This embedding model is the highest scoring⁴ model on Semantic Textual Similarity (STS) task and on average on the Massive Text Embedding Benchmark (MTEB; Muennighoff et al., 2023) making this a suitable embedding model for our purpose, which is to compare semantic similarities for measuring diversity.

Finally, I use a one-tailed paired t-test to test statistical significance for the VS_{jina} , VS_{n-gram} , and TE_{jina} results, with the hypothesis that the mean diversity of the base model is greater than that of the instruct model. In the experiment, I compare the Base (*comp*) with the Instruct (*comp*) and Instruct (*chat*). Finally, to measure the effect of the chat template, an additional test is done comparing the Instruct (*comp*) with Instruct (*chat*). In all cases, the null hypothesis is that the mean output diversity of the base model is less than the mean output diversity of the instruct model.

5.3.2 Results

Does the diversity gap exist? Figure 5.1a shows that along the VS_{jina} , VS_{n-gram} , and TE_{jina} , there is a drop in diversity for outputs generated by instruct models (for all comparisons $p < .001$ using the one-tailed t-test). The diversity gap is also reflected in the Recall scores with a clear drop. On the other hand, Precision improves (Figure 5.1b), suggesting that the generator is better at producing realistic outputs. This shows that the generated text fits well within the reference manifold. However, the decline in Recall

³See <https://github.com/krishnap25/mauve?tab=readme-ov-file#best-practices-for-mauve>

⁴On 13 January 2025. See <https://huggingface.co/spaces/mteb/leaderboard>

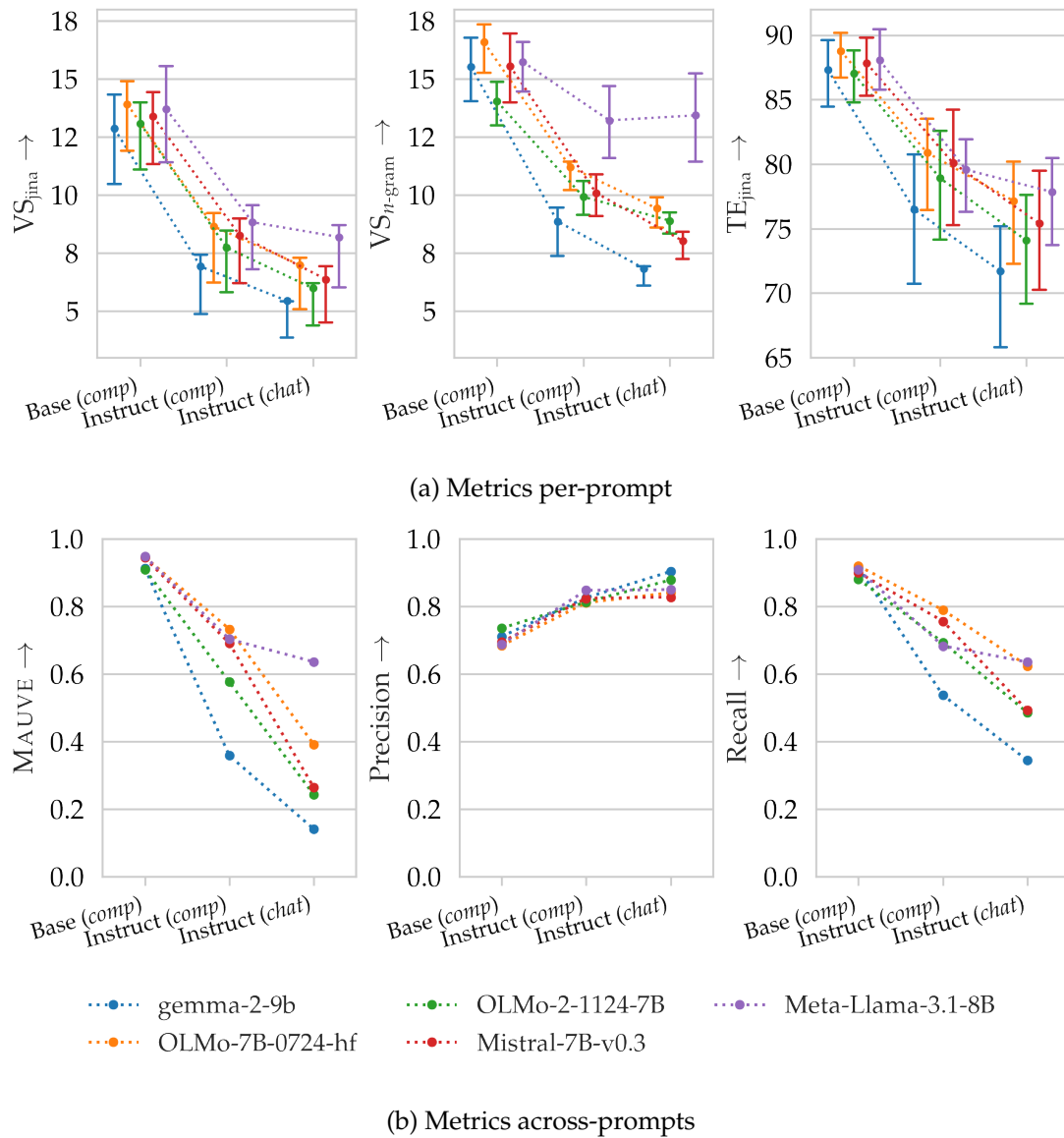


Figure 5.1: Here, the base and instruct models are compared for both prompts, *comp* and *chat*, on the narrative generation task. Figure 5.1a shows a drop in diversity for all metrics for each model ($p < .001$). The chat template typically further reduces output diversity. In Figure 5.1b, MAUVE and Recall further show that instruction tuning reduces output diversity. Precision increases, indicating a trade-off between quality and diversity due to instruction tuning. The error bars indicate the 50th percentile interval.

shows that fewer reference texts fall within the generated manifold, indicating that instruction-tuning forces a trade-off between quality and diversity. The MAUVE scores for the base models are high, demonstrating a good balance of high quality and high output diversity. This may be expected since the Writing Prompts dataset could easily be part of the training mix for any of these LLMs. However, the other diversity scores for the instruct models are much lower (Figure 5.1a), especially for the conversation prompt, suggesting that the drop for MAUVE is likely to be due to the generator producing many similar outputs. Overall, the results in Figure 5.1 clearly confirm that instruction-tuning reduces the output diversity of LLMs.

What is the impact of each instruction-tuning step? To further strengthen the findings above, the in-depth experiments for OLMo 0724 7B and OLMo 2 7B 1124 show that both SFT and DPO impact the output diversity of the model, but that the new fine-tuning step, RLVR, has little effect, as shown in Figure 5.2. This might be due to the task-specific design of RLVR (Lambert et al., 2025), which does not include narrative generation. For Recall and MAUVE, Figure 5.2 reveals that DPO has a much stronger effect than SFT, showing that reward modelling according to human preference has a particularly negative effect on LLM output diversity. Moreover, the Precision and Recall results hint that generated texts constitute a much smaller manifold than the reference due to DPO; they suggest a direction for future investigation.

What is the impact of conversation templates? Prompting the instruct models using the conversation prompt further pushes the model into a narrow output space, as it yields lower scores on diversity and quality compared to the completion prompt ($p < .01$ for all comparisons, except $VS_{n\text{-gram}}$ for Llama 3.1 8B). This is expected as the chat template is part of the training setup, and using the completion prompt on instruct models is less restrictive.

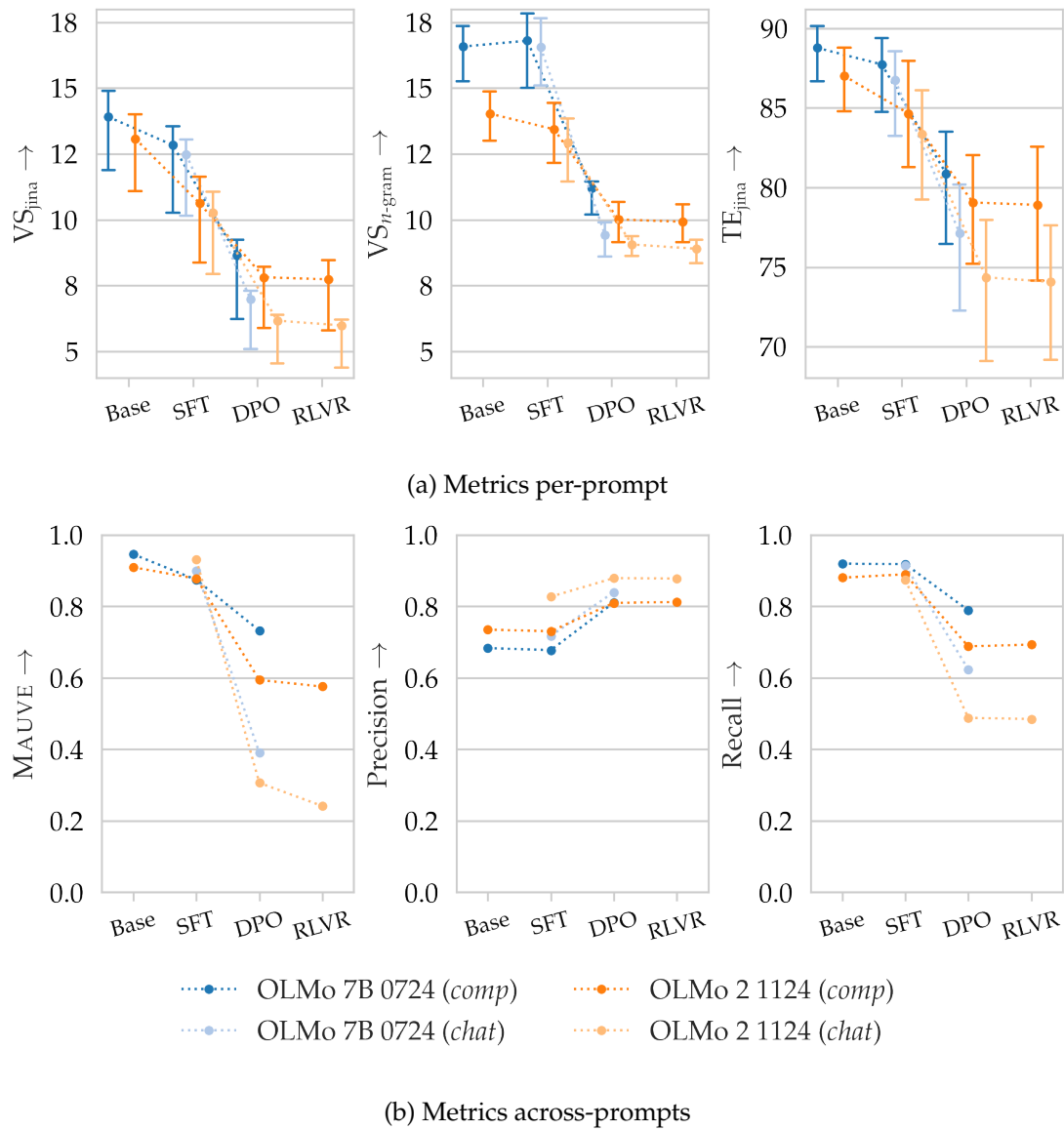


Figure 5.2: This figure presents experiments that investigate how diversity is lost with each fine-tuning step for OLMo 0724 7B and OLMo 2 7B 1124. In (a), there is an increasing decline in diversity with each fine-tuning step. In (b), there is a steep drop for the DPO models as measured by MAUVE and Recall, suggesting that reward-based fine-tuning has a particular negative effect on output diversity. The error bars indicate the 50th percentile interval.

5.4 Conformative Decoding

Motivated by my results verifying the existence of the diversity gap, I present a sampling decoding strategy that aims to reintroduce some diversity in the outputs of instruct models. The idea is that we can use base models to inform the instruct models and enable higher diversity while leveraging the usefulness and productivity of the instruction following models.

5.4.1 Decoding Strategy

Our sampling decoding strategy is a weighted sum to push the instruct model to conform to its more diverse base model (Equation 5.8). Hence, I call it *conformative decoding*. This method is related to *contrastive decoding* (Li et al., 2023; O’Brien & Lewis, 2023), but instead of using a small model to penalise bad language patterns, we use the base model to boost probabilities of tokens to yield more diverse outputs. The key insight for conformative decoding is that the base model is more diverse than instruction-tuned model as discussed in Section 5.3.2. Instead, it uses the original base model instead of ensemble methods, like a mixture-of-experts (Shazeer et al., 2017; Jiang et al., 2024) which use a (learned) selection procedure and models trained towards different tasks. It is essential that conformative decoding is used in combination with a truncation strategy, such as top- k (Fan, Lewis & Dauphin, 2018), nucleus sampling (Holtzman et al., 2020), or another truncation method to mitigate the softmax bottleneck (see Section 2.2.2 for more detail). Conformative decoding is applied to the truncated probability distribution of the instruct model. Initial explorations showed a greater increase in diversity if used without truncation, but this often results in degenerate outputs. Applying the mixture after truncation is conceptually more elegant, as it rebalances the probabilities only when the instruct model allows it. This prevents the degenerate outputs while reducing the magnitude of the conforming effect. However, note that truncation should not be too restrictive if aiming for diversity.

$$\begin{aligned} \text{ConformativeDecoding}(x_t | x_{<t}) = \\ \gamma \log p_\theta(x_t | x_{<t}) + (1 - \gamma) \log p_\phi(x_t | x_{<t}) \quad \text{for } x_t \in \mathcal{V}_{\text{valid}} \end{aligned} \quad (5.8)$$

where θ are the parameters of the instruct model, and ϕ the parameters of its base model, γ is a hyperparameter to control the mix of the two models, and $\mathcal{V}_{\text{valid}}$ is the set of valid tokens according to the chosen truncation strategy. The experiments use nucleus sampling with $p = .95$ where $\mathcal{V}_{\text{valid}}$ is the minimal set that covers at least 95% of the probability mass. The log probabilities for all $x_t \notin \mathcal{V}_{\text{valid}}$ are set to $-\infty$, meaning they are not considered in the sampling process. In practice, the mixture can be applied to the logits directly for efficiency, but this is equivalent after further softmax normalisation.

5.4.2 Experiment Setup

The method in this section is motivated by the findings regarding the diversity gap in section 5.3.2. The aim is to compare against a baseline instead of different decoding strategies. The experiments are conducted using an identical setup as described in section 5.3.1, except the aim is to see if conformative decoding improves diversity while maintaining quality. The prompts follow the conversation template and evaluate the outputs using the same metrics and, again, using the incipit so that the base model's tendencies align accurately with the task. This experiment tests two configurations of the same narrative generation task. Configuration A is the baseline using nucleus sampling ($p = .95$), and configuration B uses nucleus sampling with conformative decoding ($p = .95, \gamma = .5$). A lower γ would increase diversity, and vice versa. To test the sensitivity of the γ hyperparameter, we run a sensitivity analysis for the following values of $\gamma \in \{.1, .2, \dots, .9\}$. The chosen $\gamma = .5$ is a good and neutral value balancing quality and diversity (Figure 5.4). Finally, we test for significance using a one-tailed paired t-test with the hypothesis that the baseline has a lower mean diversity than our method.

5.4.3 Results

The result in Figure 5.3a shows that for VS_{jina} , $VS_{n\text{-gram}}$, and TE_{jina} we see diversity increases for all models ($p < .001$) except for Llama 3.1 8B. For Llama 3.1 8B, $VS_{n\text{-gram}}$ even measures a significant decline ($p < .001$), but interestingly, across-prompts, there is a recovery of diversity and even a minor increase in Precision (quality).

Looking at the across-prompt metrics, they all increase for each model (Figure 5.3b), showing that our method not only improves diversity but also maintains and even slightly increases quality, as measured by Precision. Overall, the improvements are significant but modest. However, this is not unexpected since the sampling distribution is truncated according to the instruct model. Having a less restrictive truncation strategy would naturally introduce more diversity; the effect of the mixing weight itself is examined in section 5.4.3.

What is the effect of the γ hyperparameter?

It follows from the diversity gap phenomenon (Figure 5.1) that lower values of γ should lead to higher diversity scores as more information from the base model is mixed into the distribution, and vice versa. For metrics VS_{jina} , $VS_{n\text{-gram}}$, and TE_{jina} , γ smoothly modulates the diversity gap, steadily reducing diversity as γ increases (Figure 5.4a). The relationship between Precision (quality) and Recall (diversity) averaged over models in Figure 5.4b suggests a reasonable trade-off between quality and diversity occurs for γ values around .5.

5.4.4 Human Evaluation

The automatic evaluation metrics show that conformative decoding yields more diverse outputs and even suggests a small increase in quality. To verify this in practice, I conduct a human evaluation experiment to evaluate diversity and quality using direct comparison⁵.

⁵The tasks are modelled after the human evaluation done by Finlayson et al. (2024).

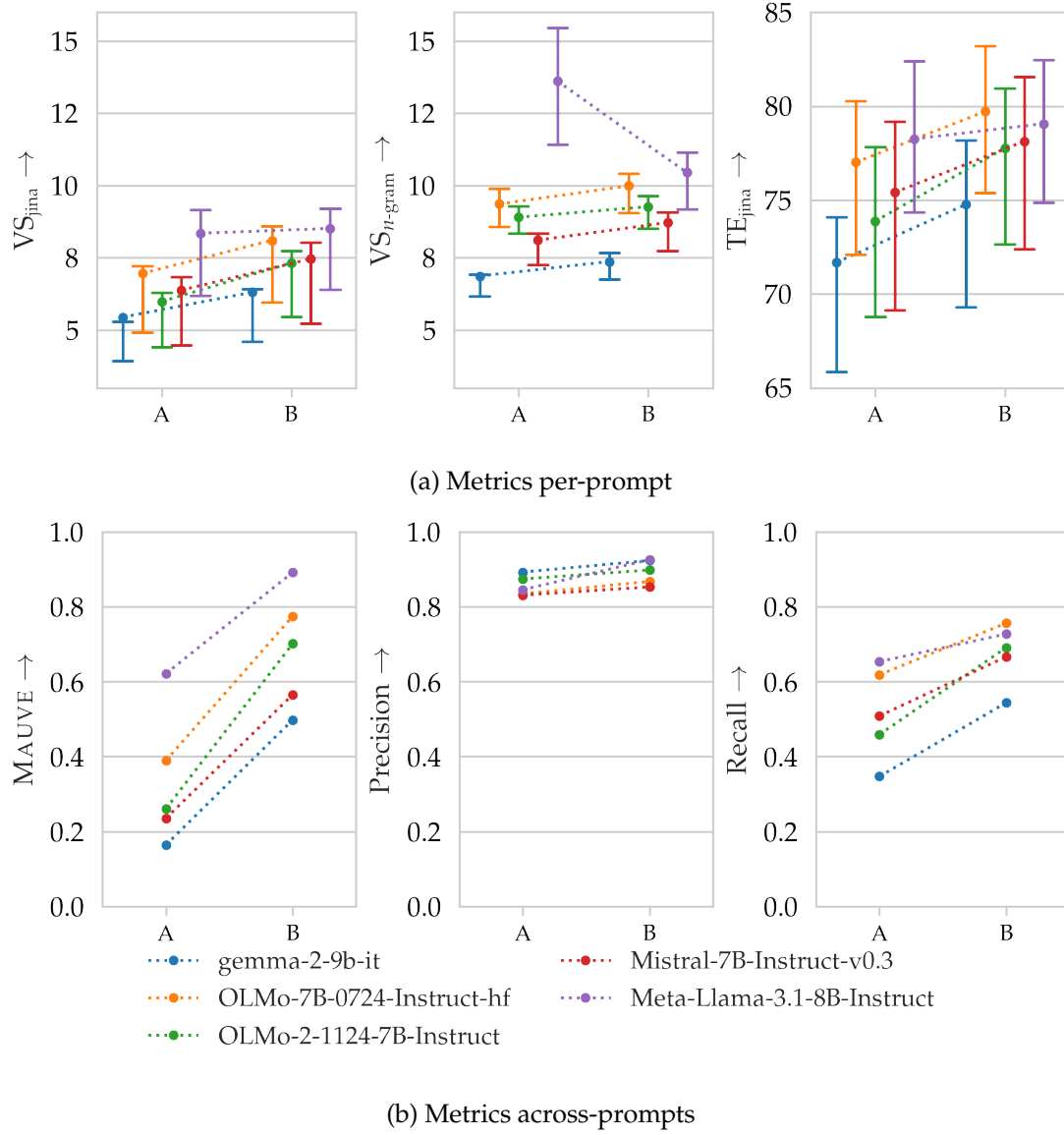
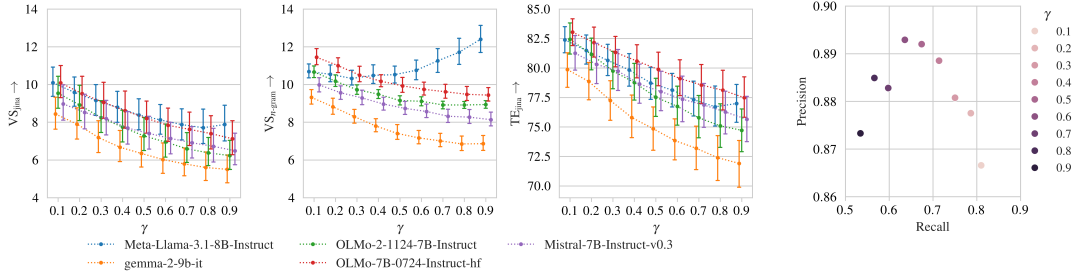


Figure 5.3: This figure shows automatic evaluations for configuration A (nucleus sampling using $p = 0.95$) and configuration B (nucleus sampling with conformative decoding using $p = 0.95$ and $\gamma = 0.5$). In (a), it shows that conformative decoding improves output diversity for all models ($p < .001$), except for Meta Llama 3.1 8B. In (b), MAUVE and Recall also show increased diversity, while Precision indicates slightly improved quality. The error bars indicate the 50th percentile interval.



(a) Diversity metrics VS_{jina} , VS_{n-gram} , and TE_{jina} with respect to γ . (b) Precision and Recall trade-off with respect to γ , averaged over models.

Figure 5.4: The effect of hyperparameter γ on conformative decoding. Figure 5.4a shows a smooth modulation of diversity as γ changes. Figure 5.4b shows a reasonable trade-off (the peak) between quality and diversity for γ values around 0.5.

Participants are asked to perform two kinds of tasks: to measure whether two sets of stories are diverse and to compare two stories to identify which is of higher quality. This experiment is done on randomly selected sets of four stories, generated by OLMo 2 7B 1124, creating a set for each prompt for both configurations, A and B, ensuring that both contain stories generated using the same four incipits. Moreover, to keep the task manageable for participants, each story was truncated to 150 tokens, and the participants were instructed not to penalise the stories for ending abruptly.

- In the diversity task, the participant is given two sets of four texts and is told their task is to select the set with higher diversity. This can mean different topics, narratives, writing styles, the depth of the story, and so on.
- In the quality task, the participant is given two similar texts and is told their task is to select the text that is of higher quality. This can mean fewer issues with grammar, coherence, cohesion, fluency, repetition, and if the narrative makes logical sense.

The participants were given pointers for both diversity and quality instead of concrete definitions or specific questions (see appendix C.1). The experiment is done via an

anonymous survey⁶ seeded through our social networks, receiving responses from 41 participants.

Human Evalaution Analysis

In total, each participant completes 4 diversity and 16 quality evaluations of stories (see examples in appendix C.3), and provides a brief explanation for each judgement. The experiment was conducted via an anonymous survey seeded through our social networks, receiving (partial) responses from 41 participants. For 31 prompts, we got the necessary 3 evaluations, allowing us to obtain a majority vote for analysis; the others were discarded. The response data are Bernoulli-distributed, so we use a binomial test to test whether the responses are due to chance.

In total, each participant evaluated four sets of stories (see examples in appendix C.3). First, they evaluated a set for diversity, followed by each story pair (matching incipits) to evaluate quality, and so on. The response data is Bernoulli distributed; hence, the responses are checked for significance if they are not the result of chance, using the binomial test. To obtain a majority vote for the analysis, each prompt required 3 responses, which was the case for 31 prompts; the others were discarded.

The results in Table 5.1 are *inconclusive* since there is no consensus for either diversity or quality, suggesting that differences indicated by the automatic evaluation metrics are too subtle for humans to detect easily, or they are not meaningful to humans. This may also be due to the lack of a precise definition, as discussed further in Section 5.5.2.

Qualitative Analysis

Since the participants provided a short justification for their judgment, I did some content analysis for all completed tasks. The initial coding was created by employing

⁶Due to it being an anonymous survey and the ethics review process we did not request demographic data. The participants likely did not have familiarity with CC, and should be considered non-expert evaluators.

Task	Prefers A	Prefers B	Total Trials	Ratio
Diversity	17	14	31	.452
Quality	62	62	124	.500

Table 5.1: Human evaluation done on the stories generated by OLMo 2 7B 1124, taking majority vote per set or text comparison. Configuration A is nucleus sampling only, and configuration B is nucleus sampling with conformative decoding. The results are not significant and inconclusive since there is no consensus.

an LLM⁷, followed by a thorough manual check and correcting issue as necessary. This yielded eleven distinct codes (Table 5.2) for both the diversity and quality tasks (see Appendix C.2 for full descriptions). The scheme and coding was checked by an external expert in qualitative analysis.

To summarise, to judge diversity, participants typically focus on the topic variety, style variety, and differences in depth and detail as part of their judgments; for quality, the focus is typically on coherence, structure, and engagement.

5.5 Discussion and Limitations

The result of the diversity gap experiments provides additional evidence to the diversity gap between base and instruct models previously identified by Hämäläinen, Tavast and Kunnari (2023), Chu, Chen and Nakayama (2024), Kirk et al. (2024) and Le Bronnec et al. (2024). Our work explores this issue explicitly in the creative domain of narrative generation. The automatic evaluation presented shows clearly that the diversity gap exists for multiple instruct LLMs. Instruction-tuning leads the model to focus on the conversation task and narrow its output space. Fine-tuning toward human preference, in other words, *how* the model responds, is a strong signal that further limits output

⁷Using Claude Fable

Diversity ($n = 146$)		Quality ($n = 579$)	
Code	Explanations	Code	Explanations
Depth / detail	42	Coherence / sense	296
Topic variety	38	Structure / flow	148
Style variety	31	Engagement	99
Quality / readability	31	Narrative content	52
Mood / tone / emotion	15	Imagery / detail	50
Structure / form	14	No reason / undecided	43
Vocabulary	13	Word choice	38
Perspective / characters	12	Pacing / length	20
Repetition / sameness	11	Grammar / mechanics	20
Similar / uncertain	11	Repetition	19
Uncodable	1	Authenticity (AI)	5

Table 5.2: Code frequencies from the basic content analysis of the participants judgment justifications for the diversity and quality tasks. Coding is multi-label. Code definitions are given in Appendix C.2.

diversity.

The method presented in this chapter relies on a truncation strategy to avoid degenerate outputs (Holtzman et al., 2020). Here, I use nucleus sampling as the baseline; it could be worthwhile to investigate other truncation strategies (e.g., Fan, Lewis & Dauphin, 2018; Meister et al., 2023; Finlayson et al., 2024) and how they interact with the choice of γ (section 5.4.3). Moreover, it might be interesting to adapt conformative decoding into an objective function for beam search, similar to contrastive decoding proposed by Li et al. (2023).

This study focuses solely on smaller open-source and open-weights models for practical reasons. Smaller models are less powerful and more likely to benefit from our decoding

strategy, but also running the base and instruct models simultaneously increases the computational overhead. However, Le Bronnec et al. (2024) showed that Llama 2 70B Chat also suffers a loss in output diversity, and future work should investigate the effects of our decoding algorithm for larger LLMs and in different domains that value diversity.

Finally, using incipits introduces some diversity, as the samples were not drawn for identical writing prompts. However, as mentioned in section 5.3.1, this is a necessary limitation for a fair comparison with the base models to make them perform the narrative generation task consistently. Omitting the incipit leads to comparing artefacts of different classes, which is not meaningful. Future work could investigate via an ablation study what the ideal length of an incipit is without altering the artefact class.

5.5.1 Why does Llama 3.1 8B behave differently?

Llama 3.1 8B is the one model that deviates from the trend in a few ways, and it is worth discussing why. For the $VS_{n\text{-gram}}$ metric, the chat template does not further reduce diversity (Figure 5.1a). One possibility is that Llama 3.1 8B’s preference tuning stage reduces diversity less aggressively than those of other models. A second possibility concerns the training data: the particular mix Llama 3.1 8B was trained on may yield higher baseline output diversity that preference tuning does not fully collapse. This is parallel to Zhou et al. (2023a), who show that Llama 1 (Touvron et al., 2023a) can be improved by fine-tuning on fewer but higher-quality and more diverse examples, suggesting that the composition of the training data may matter more than its sheer volume in shaping output diversity. The third possibility concerns training volume: Llama 3.1 8B was trained on a substantially larger budget of 15 trillion tokens⁸ than the other models (the next largest being Gemma 2 9B at 8 trillion⁹), and this broader exposure may help it retain the lexical variation $VS_{n\text{-gram}}$ is designed to capture, rather than collapsing onto a narrower set of tokens after preference tuning. These explanations are mutually confounded with each other and with architecture, so the behaviour cannot

⁸See <https://huggingface.co/meta-llama/Llama-3.1-8B>

⁹See <https://huggingface.co/google/gemma-2-9b>

be attributed to a single factor.

Conformative decoding also proved less effective for Llama 3.1 8B as measured by $VS_{n\text{-gram}}$, VS_{jina} , and TE_{jina} (Figure 5.3a). This is consistent with the picture above: it suggests the method is most suitable for models that exhibit low diversity after preference tuning, and its effectiveness may depend on the size of the diversity gap it has to close. In the across-prompts results (Figure 5.3b), we still observe increased diversity, suggesting the per-prompt evaluation is smoothed across prompts. A follow-up study should investigate the relationship between the degree of the diversity gap and the improvements due to conformative decoding.

5.5.2 Challenges for Human Evaluations of Diversity

Algorithmic metrics are useful for detecting *whether* there is a difference, but are unable to tell *what* that difference is. Human evaluations can help identify what those differences are, though our evaluation shows this is challenging because those differences may be too subtle to be easily detected. There is a broader issue of definition: while the evaluation of creativity has been extensively documented (Lamb, Brown & Clarke, 2015); less attention has been given to diversity, although recently work has started in this direction (Berns, 2024). Text diversity is task- and domain-dependent; the criteria for diversity in narratives may differ substantially from the criteria for journalistic artefacts or even poems. Text diversity might simply be too ill-defined for consistent human evaluation, as evidenced by the variety of strategies reported by participants. These observations provide ample direction for future work, but also suggest that more expertise might be required in evaluations. Current experimental methods are limited in scope and frequently fail to investigate the specific qualities that make a set of outputs diverse; consequently, it is unclear how automatic evaluation metrics align with those qualities. It is essential to develop a systematic approach towards accurately evaluating diversity.

Diversity is a tricky concept to capture for narratives, especially for longer sequences. Similarly, it is difficult to assess by human participants (Shaib et al., 2025). Assessing the

diversity of a set of narratives requires extensive reading, causing fatigue among participants. This gets further complicated if the setup is a direct comparison task. Narratives are complex objects that span multiple states and transitions (Gervás, 2009). Hence, truncating the story to an appropriate narrative length, even further than described in the experiment above, might heavily impact perceived diversity.

The tasks presented in the human evaluation experiment, without specific definitions and questions, proved to be difficult. The human evaluation experiment highlights two challenges: First, the difference in diversity between the configurations is perhaps too subtle for humans to detect. Generally speaking, richer embeddings imply better results, and while the chosen model seems appropriate, it might be that each embedding model learns towards a different kind of diversity or quality that is not easily detected by humans. Secondly, although the task included pointers for diversity and quality based on informal feedback, this was often interpreted as a definition. Since the pointers leave room for interpretation, this potentially biases the results.

Participants were recruited through personal and professional networks via an anonymous survey, which means no demographic information was collected. The participant pool should be considered non-expert, which may introduce noise into the evaluations; however, since the stories generated by the LLMs cannot be considered high literature (see Appendix C.3), using domain experts might skew the results in the opposite direction, as experts applying higher literary standards to mediocre outputs may evaluate all stories as similarly poor, reducing sensitivity to diversity differences. Nevertheless, the inconclusive result of the human evaluation may be attributable to the limited representativeness of the participant sample; the recruitment channels make it likely that the participants were predominantly Western European and university-educated, though this cannot be verified without demographic data.

There is considerable work on evaluating creativity (Lamb, Brown & Clarke, 2015), but for diversity, not so much. Diversity might still be too ill-defined, and more work is necessary to standardise criteria and (automated) scalable evaluation methods. For the narrative generation task specifically, there may be opportunities to transform the

narrative structure into a list of actions, as generated by CC systems, such as MEXICA (Pérez y Pérez & Sharples, 2001). The open questions and issues highlighted here are important across research disciplines since it is unclear how current metrics and embedding models align with human preferences in practice. Current experimental methods only get us so far and often do not investigate in what way a set of outputs is diverse.

5.6 Conclusion

This chapter explores the diversity gap on a narrative generation task for five models and proposes a decoding strategy motivated by the result to reintroduce some diversity. In the experiments, diversity is evaluated on a per-prompt basis using the Vendi Score (both lexical and semantic features) and Truncated Entropy, and across-prompts using MAUVE, Precision and Recall to evaluate quality. The results show that instruction-tuned LLMs suffer a loss in diversity compared to their base models, which is mostly due to Direct Preference Optimisation. Based on this finding, I present a decoding strategy that aims to improve diversity after instruction tuning; the novel approach introduced in this chapter, conformative decoding, typically reintroduces some diversity while maintaining precision. Although the method improves output diversity, it is unable to fully overcome the diversity gap. The human evaluation experiment shows no consensus among participants for judgments about diversity and quality. This is perhaps unsurprising, given that discrepancies between automatic metrics and human judgments are well-documented, the BLEU score being a notable example (Papineni et al., 2002; Reiter, 2018). The human evaluation experiment mainly suggests larger issues with diversity being ill-defined, and that further research is needed on how to assess diversity and how to design experiments.

Chapter 6

Conclusions

The objective of my PhD research is to deepen the understanding of the creative application of Large Language Models (LLMs) through analysis using Computational Creativity (CC) evaluation frameworks and empirical investigations. I also present a decoding strategy informed by empirical findings to improve the output diversity of LLMs. In this concluding chapter, I give my answers to the main research question through the research questions as presented in Chapter 1 section 1.2.1. Moreover, I reflect on the findings of my research and present some questions that are important for using LLMs as creativity tools or producers in their own right. Finally, I provide three directions for future research that I believe are necessary to meaningfully apply LLMs to creative tasks and overcome current limitations and gaps in empirical evaluation.

6.1 Answers to Research Questions

Research Question 1 (Chapter 3)

Can we evaluate the creativity of pure large language models using existing CC frameworks?

Chapter 3 of this dissertation presents an analysis of LLMs in general and assesses if they are able to satisfy three established CC evaluation frameworks: “Mere Generation” (Ventura, 2016), the Creative Tripod (Colton, 2008), and FACE (Colton, Charnley & Pease, 2011; Pease & Colton, 2011a). Since LLMs have found uses for a wide range of tasks, many of them creative, the primary goal of this chapter is to establish if these frameworks are useful for the creativity evaluation of LLMs. These frameworks have relied on specific criteria depending on the angle each framework takes to approach the concept of creativity. To answer this research question: pure LLMs, as defined in the introduction, can readily satisfy these frameworks, but the frameworks are unable to tell how an LLM is being creative or what is responsible for its creativity. The analysis in Chapter 3 clearly shows that these frameworks fall short, primarily due to the inability to “point out” functionality (that was previously explicitly modelled) and assess how LLMs apply given inputs. This most prominently arises in the “Mere Generation” framework, where LLMs appear to use a basic prototype algorithm, and yet, they achieve the abilities of the most advanced prototype algorithms by simulating their aims. Similarly, FACE is based on creative acts performed by systems, but this frame can only attribute ⟨C⟩, the use of concepts, to machine learning algorithms, and the transformer is no exception. Colton’s Creative Tripod is based on external perception of creativity, but LLMs are so effective at producing high-quality output that this is easily satisfied. However, Colton (2008) also relies on the process, and this leads to the same issue as above: we cannot point to any specific function that drives the creative process. While it appears the LLM can do more than its mechanisms suggest, there is no reliable way to establish, via analysing the algorithm, how it performs the necessary conditions for creativity. Satisfying CC

frameworks is not enough to establish what is responsible for the creativity of LLMs: empirical observations are necessary to tease out what enables their creative behaviours.

Research Question 2 (Chapter 4)

Is temperature the universal creativity parameter of pure large language models?

Chapter 4 presents the first empirical contribution of this dissertation. Probing LLMs to evaluate their behaviours is, in many ways, complex and requires careful consideration of the experimental setup. In this chapter, I present a novel methodology for addressing Research Question 2 by developing the idea of the exemplar object for LLMs, linking Prototype Theory (Rosch, 1973; Rosch et al., 1976) and Exemplar Theory (Medin, Altom & Murphy, 1984) from Cognitive Science to the greedy sample of any given prompt. The exemplar object serves as a baseline for comparison with other objects generated for different temperature parameter values. In the human evaluation experiment, participants were tasked with assessing the creativity of stories based on the following criteria: novelty, typicality, cohesion, and coherence. The answer to the research question is that temperature as a universal creativity parameter of LLMs is far weaker and more nuanced than suggested. The results show weak correlation with novelty and negative correlation with coherence, meaning that the results did not fulfil all conditions of creativity for this specific creative task.

Research Question 3 (Chapter 5)

Does instruction-tuning reduce the output diversity of pure large language models for a narrative generation task?

The first part of Chapter 5 presents the second empirical contribution of this dissertation. This question is aimed at establishing if a known phenomenon, a reduction in output diversity due to instruction-tuning (Hämäläinen, Tavast & Kunnari, 2023; Chu, Chen

& Nakayama, 2024; Kirk et al., 2024; Le Bronnec et al., 2024; Shypula et al., 2025), also exists for a simple narrative generation task based on the Writing Prompts dataset (Fan, Lewis & Dauphin, 2018). While the domain is the main difference with previous studies, I also evaluate the diversity gap using metrics specifically designed for diversity: Vendi Score (Friedman & Dieng, 2023) and Truncated Entropy (Ibarrola, Lawton & Grace, 2024). By using these metrics, the work expands our understanding of this phenomenon beyond other common metrics based on distinct n -gram counts (e.g., Kirk et al., 2024) or Precision and Recall (e.g., Hämäläinen, Tavast & Kunnari, 2023; Le Bronnec et al., 2024). The answer to the research question is that the diversity gap phenomenon also emerges in the narrative generation domain. Replicating this finding for a creative task suggests that the effectiveness of the LLM may be negatively impacted when asking the LLM to perform creative tasks, as diversity is considered a key component of creativity (Jordanous & Keller, 2016).

Research Question 4 (Chapter 5)

Can a pure large language model produce more diverse output, and maintain quality, by informing the instruction-tuned model using its associated base model?

The second part of Chapter 5 presents a decoding strategy motivated by the finding that the diversity gap emerges for my narrative generation task. Since the base models are more diverse than instruct models, I present a method, *Conformative Decoding*, that reintroduces information from the base models into the decoding process with the aim of making the outputs more diverse. To evaluate Conformative Decoding method, I follow the diversity gap baseline established as part of Research Question 3 and additionally, evaluate the outputs of a single model, OLMo 2 1124 7B, by human evaluation using a direct comparison task, which showed no consensus. Since this concerns only a single model, the resulting human evaluation does not generalise, and this should be addressed in future work. The computational analysis shows a modest increase in diversity using Conformative Decoding while maintaining quality across multiple models. The lack of

consensus against or in favour of Conformative Decoding may indicate that the difference in diversity is too subtle to detect, or could be due to factors impacting judgements specific to narrative generation. Research Question 4 is partially answered: from a computational perspective, Conformative Decoding does produce more diverse output, suggesting this is a promising direction. However, future work needs to address methodological questions of how to empirically evaluate diversity. In the next section, I will address the limitations of this work and set out this direction for future work.

6.2 Reflection

Generative AI is a rapidly evolving area of research, and requires substantial resources to train and test these large neural networks. As such, this state of affairs poses significant challenges for academic researchers, who aim to understand the capabilities of these models, to keep up with the state-of-the-art and conduct experiments at the necessary scale. This work is not without elements that require further reflection; therefore, in this section, I reflect on limitations, discuss the impact and consequences of the findings, and sketch ideas for future work.

6.2.1 Limitations

When doing empirical work, it is inevitable to make certain design choices along the way, and as a result, the work in this dissertation contains elements that warrant further reflection. In addition to the discussion and limitations in each of the chapters, this section focuses on the generalisability of the findings and the conclusions. These limitations provide opportunities for future work and could lead to a more nuanced and informed understanding of the findings presented in this dissertation.

Model, Prompt and Domain Choices

During the course of this research, new LLMs were released at a high frequency, each often outperforming its predecessor. In any research that evaluates the behaviours of LLMs, there is the question of which models to select. For computational evaluation, it is relatively straightforward to evaluate many models if computing resources allow, but for human evaluation, testing many models, prompts and hyperparameter settings rapidly becomes infeasible.

LLMs require significant amounts of computational resources. In Chapter 4, I chose to use Llama 2 70B Chat as it was state-of-the-art at the time, but required quantisation to run on the available hardware. Quantisation can have a significant effect on quality; even a minimal reduction in floating point precision¹ has an accuracy cost that can potentially impact empirical observation. These kinds of models and hyperparameter choices all impact the reproducibility and generalisability. In Chapter 5, I took a practical approach focusing on smaller models that could fit on the hardware available, and I emphasised open-source models (which also benefits research communities). Particularly, this latter choice is more valuable for the research communities, as opposed to using closed-sourced models like Anthropic’s Claude or OpenAI’s ChatGPT. Moreover, this means that my results only apply to smaller models, although the diversity gap has been found to exist for Llama 2 70B (Le Bronnec et al., 2024), making it likely that I would have found similar results.

The research questions of this dissertation ask about general LLM behaviours regardless of the creative domain. However, to investigate these questions, it has to be tested in some domain. Since LLMs are language models, an obvious choice seems to be narrative generation, as a domain that has a long history in CC (Gervás, 2009) and is sometimes seen as the grand challenge of creative output. Language and narratives are, after all, crucial for human cultural evolution (Tomasello, 1999). However, focusing on one domain limits the breadth of the conclusion. In other domains, the evaluated models might show

¹For more information on the quantisation effects for Llama 2 70B Chat, see <https://huggingface.co/TheBloke/Llama-2-70B-Chat-GGUF>

different results on jokes or poems, which could be better or worse, depending on the training data. Similarly, the prompt could include additional information to provoke different behaviours. Throughout the dissertation, the goal was always to minimise these effects, but some variation will always be present. Future work needs to explore other domains and models to further improve upon our current understanding of the findings.

Experimental Design and Criteria for Creativity

Human evaluation is expensive and complex for the kind of work presented here. In general, evaluating creativity is not an easy task, as exemplified by the many CC evaluation frameworks in existence. For human studies, the relevant questions are which criteria to choose, and how to formulate the questions posed to participants. It is important to make a case for what criteria to use (Jordanous, 2012), but it is not straightforward what questions to ask and how to present those criteria to the participants (Lamb, Brown & Clarke, 2018). This became especially apparent through the human evaluation in Chapter 5. It may initially appear preferable to let the participant decide what diversity means, providing only pointers, but this task proved to be complex. Especially with narratives, there are many possible dimensions along which one could consider diversity. However, this kind of limitation opens up an interesting avenue for future work, as I will discuss below, since it shows that in the CC community, we have clear ideas about what we expect from a creative system, but how we ask to test certain properties is much less defined. For generative AI in particular, this seems important, as the conclusion from Chapter 3 tells us that we cannot rely on the model's process for assessing the creativity of these models.

Another important factor is the sample sizes available for these empirical experiments, as they are by no means large, and participants are primarily from a Western European demographic, most of them pursuing or in possession of university degrees. As such, it is clear that the findings in this dissertation could require a more heterogeneous sample to improve generalisability. In Chapter 4, I opted to evaluate just one model for this reason, as we were only able to recruit participants from the Computational Creativity

module at the University of Kent. For the human evaluation experiment in Chapter 5, focusing on evaluating the decoding strategy on a single model, the participants were sourced through my and my collaborators' social networks. The participants, while some were students on the Computational Creativity module, are not domain experts. Although evaluation using domain experts is generally essential (Lamb, Brown & Clarke, 2018), this is also costly and difficult to do at a scale. However, as is clear from the texts in appendices B.2, C.3, and C.4, these models do not generate anything too sophisticated, meaning evaluation using non-experts is still informative.

With limited access to participants, it is difficult to get sufficient sample sizes to test each experiment for multiple models. Computational methods can be effective in mitigating some of these limitations, providing a general qualitative impression. Especially, the effectiveness of computational methods which are specifically designed for assessing diversity (Vendi Score and Truncated Entropy), generally strengthens the analysis and conclusions, as these methods have been assessed for their effectiveness (Friedman & Dieng, 2023; Ibarrola, Lawton & Grace, 2024). Nevertheless, it is important to keep in mind that these methodologies differ. Computational methods allow computer scientists to detect differences, while human evaluation detects whether that difference matters in practice. Both are valuable, and should be used when needed.

6.2.2 Discussion

The technical discussion for each contribution are part the respective chapters. In this section, I will discuss a broader view regarding the creative application of Large Language Models and how it relates to CC.

Is evaluating pure LLMs useful or interesting?

When I started this doctoral project in January 2022, LLMs had not yet attracted widespread attention. GPT-2 (Radford et al., 2018) existed, and it was excellent for fine-tuning towards specific (creative) tasks (e.g., Sawicki et al., 2022). While more effective com-

pared to other deep learning methods, it was not yet the breakthrough that ChatGPT was. LLM-based systems today are no longer just transformers; they are augmented in various ways that enhance their capabilities beyond the discussion in Chapter 3. The research findings of this dissertation focus specifically on the evaluation of pure LLMs, but the results do not indicate any significant creative behaviours. The artefacts created by the investigated LLMs are mediocre and are mostly interesting from a research perspective. While the work here sheds light on the capabilities of pure LLMs, it appears that a more fruitful direction is to view the LLM as a component that can address complex tasks through natural language as part of a larger system.

The breakthrough of language models and multimodality has enabled the design of intricate systems that can prompt computers to generate novel approaches. The primary example is AlphaEvolve (Novikov et al., 2025), which uses an evolutionary algorithm to evolve a population of LLM-generated ideas, with their fitness judged by a larger LLM, ultimately generating novel ideas. This approach has already made progress in identifying more efficient ways to multiply matrices of specific dimensions, which is not insignificant since this is the fundamental operation for training neural networks. This method is in large part successful because it generates code, which is executable and testable. Creative domains might be more difficult to automate this way. AlphaEvolve, Generative Agents (Park et al., 2023), and augmentations like FunSearch (Romera-Paredes et al., 2024) are clearly techniques which augment LLMs and can potentially lead to more interesting creative behaviours than sampling alone would.

Why did the “temperature paper” strike a chord?

The paper that is the basis for chapter 4 was positively received. The work itself is a small and focused contribution, and is an attempt at addressing a simple question. Researchers across many disciplines have shown interest in LLMs, considering their potential applications and the role of the temperature parameter. Consequently, the success of this paper is largely attributed to its timely nature. A common narrative was that allowing more randomness in the generation process must enable creativity (Boden,

1992; Grace & Maher, 2014). While some argue that randomness is sufficient for creative behaviour (Simonton, 1999, 2023), others contend that it requires more (Ventura, 2016), and Chapter 4 showed that for an unaugmented LLM, randomness may not be sufficient. Randomness is necessary, but the greedy, or high-probability, outputs are often repetitive and lack variation (Meister et al., 2023), and in general, are uninteresting (Mondol & Brown, 2021b); this is not enough for creativity. This paper is one that addresses this simple yet common line of thinking. Another reason for the success might be that this paper seems to strike a balance between a feasible experiment and generalisability. There was no methodology to test empirically how to approach this experiment, which further complicates the issues with probing LLMs and the limitations mentioned above and in Chapter 4. It is exemplary in showing that small-scale experiments can be meaningful. It will be valuable to determine more precisely what creativity for LLMs entails, as they are generally used as tools for supporting humans to work more productively.

What might generative AI mean for Computational Creativity?

Recent models are improving at a steady pace, leading to predictions and evidence that entry-level white-collar jobs are being eliminated (Roose, 2025; VandeHei & Allen, 2025). This dissertation began with an examination of what CC is and what its challenges are. Given the substantial progress in the range of tasks that can be automated, I argue that, as a field, it is necessary to reassess these challenges. The field has inadvertently expanded into the area of Creative AI, which is gaining traction at major AI conferences².

A key difference between the traditional CC system and generative AI is that the former were typically employed as an extension of their creators. Perhaps LLMs and diffusion models are no different, but they are, in fact, products that are used by consumers. For example, AARON was continuously iterated upon as its creator, Harold Cohen, added more rules (Boden, 1992). Similarly, Tony Veale has continuously improved upon the abilities of Excalibur, eventually using an LLM as part of its generation process

²IJCAI has the AI, Arts and Creativity Track (IJCAI, 2024) and since 2023 NeurIPS promoted its workshops and showcases into a full Creative AI track (NeurIPS, 2023).

(Veale, 2024). While these two systems can autonomously produce drawings and comics, respectively, it is clear they are improved as their developers see fit; it is an extension of their personal creative output. While Computational Creativity as a field is focused on developing systems that are creative as judged by an unbiased observer (Colton & Wiggins, 2012), I believe it is reasonable to question what these systems ultimately are. The papers about those systems are blueprints for others to develop similar systems with a specific purpose, but I am no longer certain that their authors intend them to exist on their own. This is more evident because CC systems are usually closed-source developed by academics (e.g., MEXICA Pérez y Pérez and Sharples, 2001 or Excelsior Veale, 2022). LLMs are, in some sense, reshaping the kind of creativity to which we aspire. Generative, or Creative AI, as a field, is expanding rapidly, with venture capital driving research now that funders perceive AGI as within reach (Bubeck et al., 2023). Some argue that we now have everything in place to achieve open-endedness to go even further and achieve Artificial Superhuman Intelligence (Hughes et al., 2024). Although there is much debate on the exact definition of AGI (Morris et al., 2024) and whether it can be reached through scaling (Marcus, 2025), the fact that entry-level white-collar jobs are being displaced suggests that jobs involving creativity may be next, as exemplified by AlphaEvolve (Novikov et al., 2025). The more compelling creative use-cases of LLMs and generative AI appear to lie in generating ideas, or snippets that inspire (McCormack et al., 2024, e.g.), rather than full, complete artistic and musical artefacts. There is much potential for LLMs when integrated in conversational interfaces to aid in ideation and creative exploration (Sood, Llano & McCormack, 2025).

The world is being flooded with AI-generated imagery and text, which is then independently published and sold on Amazon, printed via online webshops or promoted in playlists on music streaming platforms. These works appear largely without value. Humans need to spend time creating, but also appreciating ideas, art, and music; however, where to direct that time is an increasingly challenging choice. Simply prompting generative models alone does not seem the way to create significant works of art (McCormack et al., 2023). In the next section, I call for a return to investigate the social processes and interaction involved in creativity. I argue that a single LLM (or generative AI model in

general) without augmentations is unlikely to produce anything novel and valuable on its own. The stories generated for the experiments in this dissertation are by no means impressive (see Appendices B.2 and C.4). Some artworks, due to their concepts, can only be created once or by a specific artist; when others repeat the methodology, it amounts to mere imitation. For example, only Jackson Pollock can create Jackson Pollock paintings, and anyone who attempts to place a urinal in a gallery is simply copying Duchamp. Similarly, only once can an AI-generated artwork enter and win a fine-art competition, and make a global impact³. That said, generative AI is here to stay, and as exemplified by AlphaEvolve, there is still considerable creative potential yet to be explored.

6.3 Have Large Language Models changed?

The development of and research on LLMs have made incredible leaps forward since the release of GPT-2 (Radford et al., 2019). Since conducting the research for the original paper on which this chapter is based, much has changed. Many things that LLMs could do that were surprising at the time are now normal. Much research is done that investigates “Can LLMs do X?”, but there is also an increasing body of empirical work that shows some parallels with human language learning (Millière, 2024), alignment with human judgements (Dillion et al., 2023), and even correlations between LLM text embeddings and brain activity upon observing the same text (Ren et al., 2025). However, LLMs, fundamentally, have not changed much; they are still transformer-based models based on the original work by Vaswani et al. (2017), except for activation functions and other tweaks in hyperparameters and modifications for training stability and downstream performance (Team OLMo, 2025). Due to scaling of both architecture and data, better fine-tuning approaches, multi-modality, API integration, Retrieval Augmented Generation (Lewis et al., 2020), and so on, these models have become more advanced and productive. This chapter is a reflection and response to statements and claims made in the time around the release of ChatGPT (GPT-3.5) and GPT-4, so at the end of this chapter, I conclude with

³Théâtre D’Opéra Spatial by Jason M. Allen, 2022. See <https://www.jasonmallen.com/space-opera-theater>

a commentary discussing how LLMs have progressed since, and the implications of this progress on the questions of this chapter. Additionally, in the commentary, I will further reflect on the implications for LLMs applied toward creative tasks and the consequences for CC evaluation frameworks.

6.3.1 Future Work

Aligning Criteria for Creativity with Human Evaluation Methodologies

The analysis in Chapter 3, and particularly, the human evaluation experiment in chapter 5, suggests that the current frameworks and criteria do not necessarily allow accurate evaluation of LLMs. A key issue might be that the computational methods do not align well with human evaluation methods despite being designed to do so. For example, the human evaluation experiment in Chapter 5 suggests the question of whether a human evaluation measures the same kind of diversity as the metrics, or even if humans can detect the required subtle differences in text. Computational evaluation of generative AI outputs, as discussed in chapter 5, typically involves comparing model outputs using distributional metrics such as relative entropy (Pillutla et al., 2021) and Precision & Recall (Sajjadi et al., 2018). However, since these require reference data that exhibits the desired criteria, it is crucial to validate that the intended property (e.g. diversity) is captured in that data. Alternatively, there exist entropy-based metrics that measure diversity and do not require reference data (e.g. Friedman & Dieng, 2023; Ibarrola & Grace, 2024). Computational metrics are the first step in evaluating LLMs at scale, allowing detection of initial differences in diversity. However, as discussed, it is difficult to establish that a difference in diversity matters in practice through human evaluation, especially for longer texts (Shaib et al., 2025). Human evaluation is costly, and it remains necessary to validate computational results and ensure the ground truth aligns with the desired criteria. It is essential for creativity evaluation in CC to reassess the criteria for creativity, their meaning, and how we empirically evaluate in the context of generative AI and CC models. One direction might be to investigate specific patterns in narratives more deeply through human evaluation, and analyse how these patterns affect the computational

metrics, following an approach similar to the work by Gooding and Grefenstette (2025), but applied to conversation traces.

Verifying Creative Behaviours

The previous section calls for making creativity evaluation methodologies more effective and practical; alternatively, there is room to explore interaction within the generation process. This research direction is primarily inspired by conditional design⁴, a collaboration where humans follow rules and conditions to produce novel and interesting artefacts.

Formal verification methods could be interesting to explore, as a way to guide creative behaviours the agent by imposing constraints. For instance, Finlayson et al. (2024) proposed the Basis Aware Threshold (BAT) decoding strategy, which verifies token choices against the true distribution (i.e. when the LLM is perfectly optimised) using linear programming (LP; Fletcher, 2000), which optimises an objective function given a set of linear constraints. This method is for general language use, but its approach is novel, using the softmax matrix of the LLM in the constraints. The key insight is that it can potentially be applied to multi-agent systems. Consider a main generator, a state-of-the-art LLM, and several (LLM) agents that each are tuned to attend to some specific criteria. As the main generator produces the next-token probability distributions, the goal is to verify token choices using BAT sampling with respect to the other neural agents. Each agent could be another generative neural network with a special purpose that models some neural appreciation function (Ventura, 2019), a specific style (Sawicki et al., 2023b), or certain types of documents. Investigating whether these kinds of methods can influence and guide the behaviours of LLMs applied to open-ended text generation in the domain of narrative generation can be a powerful research direction. It could potentially mitigate some of the issues that currently exist for pure transformers, such as generating artefacts of the wrong artefact class or about undesired topics.

⁴See <https://conditionaldesign.org/>

Beyond Pure Transformers: Social Creativity

From the discussion above, it is clear that moving beyond *pure transformer* is potentially more fruitful to perform creative tasks meaningfully; indeed, this migration has already happened. This involves iteratively generating and improving LLM outputs or letting LLM-powered agents interact with each other, implying a social process amongst AI agents. CC, as a field, should refresh the traditional approach of modelling cognitive functions, for example, text interpretation, reflection and text production (Hayes, 1996) for narrative generation, to induce creative behaviour in machines, and focus more on the social aspects of creativity. The community already has solid foundations on computational social creativity (CSC; Saunders & Bown, 2015) building upon work by Csikszentmihalyi (1988) and Bown (2009, 2012). In the broader creativity research community, a common narrative today is to view the nature of creativity as an assemblage that depends on different elements from an environment and not just the creativity of a human, such as the materials used in the creative process (Manta, 2018). Generative AI could be one of those elements that integrate into our creative processes. While this work predates LLMs and Generative AI, there is much to explore in adopting these ideas with the current state-of-the-art. The systems view of creativity (Csikszentmihalyi, 1988) is relevant again as a blueprint for building these multi-agent LLM systems. While systems such as FunSearch (Romera-Paredes et al., 2024) and AlphaEvolve (Novikov et al., 2025) might already map onto this framework, there is still room to further explore and understand how to design these interactions for the purpose of creative behaviour; for example, when is iteration on current ideas important? When to pass an artefact on for judgement? What is useful to share with other LLMs? Both FunSearch and AlphaEvolve generate code which is easy to evaluate, as mentioned, but CSC and its background can inform how we might make this work for more creative tasks. For example, taking inspiration from a conceptual model of reviewing and critique (Brown & Peeperkorn, 2023) might be helpful to reason about such interaction and for developing LLMs tuned to the domain of art critics or reviewers more generally. This call for renewing interest in “social interactions” between LLM-powered CC systems aligns with the initial vision to study Computational Social Creativity for this doctoral project (Peeperkorn, 2022); in

that sense, this dissertation has come full circle.

Bibliography

- Ackerman, M., & Brown, D. (2024). Depictions of jews in large generative models. In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 342–347). Association for Computational Creativity.
- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1), 147–169. [https://doi.org/10.1016/S0364-0213\(85\)80012-4](https://doi.org/10.1016/S0364-0213(85)80012-4)
- Acovino, V., Kelly, M. L., & Abdullah, H. (2023). A sci-fi magazine has cut off submissions after a flood of AI-generated stories. *NPR*. Retrieved April 24, 2023, from <https://www.npr.org/2023/02/24/1159286436/ai-chatbot-chatgpt-magazine-clarkesworld-artificial-intelligence>
- Adamopoulos, P., & Tuzhilin, A. (2014). On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology*, 5(4). <https://doi.org/10.1145/2559952>
- Agirrezabal, M., & Gonalo Oliveira, H. (2025). Refining metrical constraints in llm-generated poetry with feedback. In *Proceedings of the 16th international conference on computational creativity*. Association for Computational Creativity.

- Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, A., Qingqing andd Jansen, Roberts, A., Tagliasacchi, M., et al. (2023). MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.
- Aguilar, W., & Pérez y Pérez, R. (2015). Dev E-R: A computational model of early cognitive development as a creative process. *Cognitive Systems Research*, 33, 17–41. <https://doi.org/10.1016/j.cogsys.2014.09.002>
- Akintunde, M. E., Kevorchian, A., Lomuscio, A., & Pirovano, E. (2019). Verification of RNN-based neural agent-environment systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 6006–6013. <https://doi.org/10.1609/aaai.v33i01.33016006>
- Almeida, L. (2025, June). Uk entry-level jobs chatgpt launch adzuna. Retrieved June 30, 2025, from <https://www.theguardian.com/business/2025/jun/30/uk-entry-level-jobs-chatgpt-launch-adzuna>
- Altmeyer, P., Demetriou, A. M., Bartlett, A., & Liem, C. C. S. (2024). Position: Stop making unscientific agi performance claims. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett & F. Berkenkamp (Eds.), *Proceedings of the 41st international conference on machine learning* (pp. 1222–1242, Vol. 235). PMLR. <https://proceedings.mlr.press/v235/altmeyer24a.html>
- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43, 997–1013.
- Ammanabrolu, P., Tien, E., Cheung, W., Luo, Z., Ma, W., Martin, L., & Riedl, M. (2019a). Guided neural language generation for automated storytelling. *Second Workshop on Storytelling*, 46–55. <https://doi.org/10.18653/v1/W19-3405>
- Ammanabrolu, P., Tien, E., Cheung, W., Luo, Z., Ma, W., Martin, L., & Riedl, M. (2019b). Guided neural language generation for automated storytelling. *Second Workshop on Storytelling*, 46–55. <https://doi.org/10.18653/v1/W19-3405>

- Angluin, D. (1992). Computational learning theory: Survey and selected bibliography. *Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing*, 351–369. <https://doi.org/10.1145/129712.129746>
- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. <https://arxiv.org/abs/1607.06450>
- Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In C. S. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65–77). IGI Global. <https://doi.org/10.4018/978-1-60566-667-9.ch004>
- Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1), 32–68. Retrieved February 27, 2024, from <http://www.jstor.org/stable/416793>
- Barron, F. (1955). The disposition toward originality. *The Journal of Abnormal and Social Psychology*, 51(3), 478.
- Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A., & Criminisi, A. (2016). Measuring neural net robustness with constraints. *Advances in Neural Information Processing Systems*, 29.
- Basu, S., Ramachandran, G. S., Keskar, N. S., & Varshney, L. R. (2021). Mirostat: A neural text decoding algorithm that directly controls perplexity. *International Conference on Learning Representations*.
- Bedau, M. A., McCaskill, J. S., Packard, N. H., Rasmussen, S., Adami, C., Green, D. G., Ikegami, T., Kaneko, K., & Ray, T. S. (2000). Open problems in artificial life. *Artificial Life*, 6(4), 363–376. <https://doi.org/10.1162/106454600300103683>
- Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., Bengio, Y., & Jerbi, K. (2024). Divergent creativity in humans and large language models. <https://arxiv.org/abs/2405.13012>

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.
- Berariu, T., Czarnecki, W., De, S., Bornschein, J., Smith, S., Pascanu, R., & Clopath, C. (2023). A study on the plasticity of neural networks. <https://arxiv.org/abs/2106.00042>
- Berns, S. (2024). *Diversity in generative machine learning to enhance creative applications* [Doctoral dissertation, Queen Mary University of London]. <https://qmro.qmul.ac.uk/xmlui/handle/123456789/105581>
- Berns, S., Colton, S., & Guckelsberger, C. (2023). Towards mode balancing of generative models via diversity weights. *14th International Conference on Computational Creativity (ICCC'23)*. <https://computationalcreativity.net/iccc23/>
- Birhane, A., & Raji, D. (2022, December). ChatGPT, Galactica, and the progress trap. Retrieved April 24, 2023, from <https://www.wired.com/story/large-language-models-critique>
- Birkhoff, G. D. (1933). *Aesthetic measure*. Harvard University Press.
- Bishop, J. M., & Al-Rifaie, M. M. (2016). Autopoiesis in creativity and art. *Proceedings of the 3rd International Symposium on Movement and Computing*, 1–6.
- Boden, M. (1992). *The creative mind*. Abacus.
- Boden, M. (2004). *The creative mind: Myths and mechanisms*. Routledge.
- Bodily, P. M., & Ventura, D. (2024a). Operationalizing essential characteristics of creativity in a computational system for music composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1), 447–455. <https://doi.org/10.1609/aaai.v38i1.27799>
- Bodily, P. M., & Ventura, D. (2024b). Operationalizing Essential Characteristics of Creativity in a Computational System for Music Composition. *Proceedings of the AAAI*

- Conference on Artificial Intelligence*, 38(1), 447–455. <https://doi.org/10.1609/aaai.v38i1.27799>
- Bommarito, M. J., & Katz, D. M. (2022). GPT takes the bar exam. *arXiv preprint arXiv:2212.14402*.
- Bopp, J. A., Vornhagen, J. B., & Mekler, E. D. (2021). "my soul got a little bit cleaner": Art experience in videogames. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY). <https://doi.org/10.1145/3474664>
- Bowman, S. R. (2023). Eight things to know about large language models. <https://arxiv.org/abs/2304.00612>
- Bown, O. (2009). Against individual creativity. *Dagstuhl seminar proceedings*.
- Bown, O. (2012). Generative and adaptive creativity: A unified approach to creativity in nature, humans and machines. In *Computers and creativity* (pp. 361–381). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31727-9_14
- Bown, O. (2021). *Beyond the creative species: Making machines that make art and music*. The MIT Press. <https://doi.org/10.7551/mitpress/10913.001.0001>
- Bradshaw, G. F., Langley, P. W., & Simon, H. A. (1983). Studying scientific discovery by computer simulation. *Science*, 222(4627), 971–975. <https://doi.org/10.1126/science.222.4627.971>
- Brejová, B., Brown, D. G., Li, M., & Vinař, T. (2005). ExonHunter: A comprehensive approach to gene finding. *Bioinformatics*, 21 suppl 1, i57–i65. <https://doi.org/10.1093/bioinformatics/bti1040>
- Breum, S. M., Egdal, D. V., Mortensen, V. G., Møller, A. G., & Aiello, L. M. (2024). The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18, 152–163. <https://doi.org/10.1609/icwsm.v18i1.31304>
- Brooks, F. P. (1996). The computer scientist as toolsmith II. *Commun. ACM*, 39(3), 61–68. <https://doi.org/10.1145/227234.227243>

- Brown, D. G., & Jordanous, A. (2022). Is style reproduction a computational creativity task? In M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz & T. Veale (Eds.), *Proceedings of the 13th international conference on computational creativity* (pp. 220–229). Association of Computational Creativity.
- Brown, D. G., & Mondol, T. (2021). On the problem of small objects. *Entropy*, 23(11). <https://doi.org/10.3390/e23111524>
- Brown, D. G., & Peeperkorn, M. (2023). Reviewing, creativity, and algorithmic information theory. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 133–147).
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Bryan-Kinns, N., Banar, B., Ford, C., Reed, C. N., & Armitage, J. (2024). Explainable AI and Music. In Y. Zhang & L. Mou (Eds.), *Artificial Intelligence for Art Creation and Understanding* (1st). CRC Press. <https://doi.org/10.1201/9781003406273>
- Bryan-Kinns, N., Banar, B., Ford, C., Reed, C. N., Zhang, Y., Colton, S., & Armitage, J. (2021). Exploring XAI for the arts: Explaining latent space in generative music. *eXplainable AI Approaches for Debugging and Diagnosis*.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. <https://arxiv.org/abs/2303.12712>
- Bundy, A. (1994). What is the difference between real creativity and mere novelty? *Behavioral and Brain Sciences*, 17(3), 533–534.
- Calderwood, A., Qiu, V., Gero, K. I., & Chilton, L. B. (2020). How novelists use generative language models: An exploratory user study. *HAI-GEN+user2agent*.

- Carlson, K., Riddell, A., & Rockmore, D. (2018). Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10).
- Casebourne, I. (1996). The grandmother program: A hybrid system for automated story generation. *Second International Symposium of Creativity and Cognition (Loughborough, England, 1996)*, 146–155.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3). <https://doi.org/10.1145/1541880.1541882>
- Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for bayesian cognitive science*. Oxford University Press, USA.
- Chen, H., & Ding, N. (2023). Probing the “creativity” of large language models: Can models produce divergent semantic association? *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12881–12888. <https://doi.org/10.18653/v1/2023.findings-emnlp.858>
- Chen, J., Gao, C., Yuan, S., Liu, S., Cai, Q., & Jiang, P. (2025). DLCRec: A novel approach for managing diversity in LLM-based recommender systems. <https://arxiv.org/abs/2408.12470>
- Cherry, E., & Latulipe, C. (2014). Quantifying the creativity support of digital tools through the creativity support index. *ACM Trans. Comput.-Hum. Interact.*, 21(4). <https://doi.org/10.1145/2617588>
- Chiang, T. (2023). ChatGPT is a blurry JPEG of the web. *The New Yorker*. Retrieved February 6, 2023, from <https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>
- Chomsky, N. (1965, May). *Aspects of the theory of syntax*. MIT Press.
- Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4302–4310.

- Chu, K., Chen, Y.-P., & Nakayama, H. (2024). Exploring and controlling diversity in LLM-agent conversation. <https://arxiv.org/abs/2412.21102>
- Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology*, 6. <https://doi.org/10.33011/lilt.v6i.1245>
- Cilibrasi, R., Vitányi, P., & de Wolf, R. (2004). Algorithmic clustering of music based on string compression. *Computer Music Journal*, 28(4), 49–67.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition*.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. <https://arxiv.org/abs/2110.14168>
- Cohen, N. (2007). Google halts ‘miserable failure’ link to president bush. *The New York Times*.
- Colton, S. (2002). The HR program for theorem generation. In A. Voronkov (Ed.), *Automated deduction—cade-18* (pp. 285–289). Springer Berlin Heidelberg.
- Colton, S. (2008). Creativity versus the perception of creativity in computational systems. *2008 AAAI Spring Symposium: Creative Intelligent Systems*.
- Colton, S. (2012). The painting fool: Stories from building an automated painter. In J. McCormack & M. d’Inverno (Eds.), *Computers and creativity* (pp. 3–38). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31727-9_1
- Colton, S., Charnley, J., & Pease, A. (2011). Computational creativity theory: The FACE and IDEA descriptive models. In D. Ventura, P. Gervás, D. F. Harrell, M. L. Maher, A. Pease & G. Wiggins (Eds.), *Proceedings of the second international conference on computational creativity* (pp. 90–95).

- Colton, S., Goodwin, J., & Veale, T. (2012). Full-FACE poetry generation. In M. L. Maher, K. Hammond, A. Pease, Rafael Pérez, D. Ventura & G. Wiggins (Eds.), *Proceedings of the third international conference on computational creativity* (pp. 95–102).
- Colton, S., & Muggleton, S. (2006). Mathematical applications of inductive logic programming. *Machine Learning*, 64(1), 25–64. <https://doi.org/10.1007/s10994-006-8259-x>
- Colton, S., Pease, A., Corneli, J., Cook, M., Hepworth, R., & Ventura, D. (2015). Stakeholder groups in computational creativity research and practice. In T. R. Besold, M. Schorlemmer & A. Smaill (Eds.), *Computational creativity research: Towards creative machines* (pp. 3–36). Atlantis Press. https://doi.org/10.2991/978-94-6239-085-0_1
- Colton, S., Pease, A., Guckelsberger, C., McCormack, J., & Llano, T. (2020). On the machine condition and its creative expression. In T. V. F. Amílcar Cardoso Penousal Machado & J. M. Cunha (Eds.), *Proceedings of the 11th international conference on computational creativity* (pp. 342–349). Association for Computational Creativity.
- Colton, S., Pease, A., & Saunders, R. (2018). Issues of authenticity in autonomously creative systems. In F. Pachet, A. Jordanous & C. León (Eds.), *Proceedings of the 9th international conference on computational creativity* (pp. 272–279). Association for Computational Creativity.
- Colton, S., & Wiggins, G. A. (2012). Computational creativity: The final frontier? *Proceedings of the 20th European Conference on Artificial Intelligence*, 21–26.
- Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1), 51–73. <https://doi.org/10.1080/09298219508570672>
- Cook, M., & Colton, S. (2018). Redesigning computationally creative systems for continuous creation. In F. Pachet, A. Jordanous & C. León (Eds.), *Proceedings of the 9th international conference on computational creativity* (pp. 32–39). Association for Computational Creativity.
- Cook, M., Colton, S., & Pease, A. (2012). Aesthetic Considerations for Automated Platform Design. *Proceedings of the AAAI Conference on Artificial Intelligence and*

- Interactive Digital Entertainment*, 8(1), 124–129. <https://doi.org/10.1609/aiide.v8i1.12524>
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd). Wiley.
- Csikszentmihalyi, M. (1988). Society, culture, and person: A systems view of creativity. In *The nature of creativity* (pp. 243–270). Cambridge University Press.
- Csikszentmihalyi, M. (2014). Society, culture, and person: A systems view of creativity. In *The systems model of creativity*. Springer Netherlands. https://doi.org/10.1007/978-94-017-9085-7_4
- Das, P., & Varshney, L. R. (2022). Explaining artificial intelligence generation and creativity: Human interpretability for novel ideas and artifacts. *IEEE Signal Processing Magazine*, 39(4), 85–95. <https://doi.org/10.1109/MSP.2022.3141365>
- DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., ... Zhang, Z. (2025a). Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning. <https://arxiv.org/abs/2501.12948>
- DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., ... Pan, Z. (2025b). Deepseek-v3 technical report. <https://arxiv.org/abs/2412.19437>
- DeMers, D., & Cottrell, G. (1992). Non-linear dimensionality reduction. In S. Hanson, J. Cowan & C. Giles (Eds.), *Advances in neural information processing systems* (Vol. 5). Morgan-Kaufmann.
- den Heijer, E., & Eiben, A. E. (2010). Comparing aesthetic measures for evolutionary art. *Applications of Evolutionary Computation*, 311–320.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>

- Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2022). 8-bit optimizers via block-wise quantization. *International Conference on Learning Representations*. <https://openreview.net/forum?id=shpkpVXzo3h>
- Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7), 597–600. <https://doi.org/https://doi.org/10.1016/j.tics.2023.04.008>
- Dohare, S., Hernandez-Garcia, J. F., Lan, Q., Rahman, P., Mahmood, A. R., & Sutton, R. S. (2024). Loss of plasticity in deep continual learning. *Nature*, 632(8026), 768–774. <https://doi.org/10.1038/s41586-024-07711-7>
- Dollinger, S. J., Urban, K. K., & James, T. A. (2004). Creativity and openness: Further validation of two creative product measures. *Creativity Research Journal*, 16(1), 35–47. https://doi.org/10.1207/s15326934crj1601_4
- Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. <https://doi.org/10.1126/sciadv.adn5290>
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.
- D'Souza, R. (2021). What characterises creativity in narrative writing, and how do we assess it? research findings from a systematic literature search. *Thinking Skills and Creativity*, 42, 100949. <https://doi.org/https://doi.org/10.1016/j.tsc.2021.100949>
- Du, W., Yang, Y., & Welleck, S. (2025). Optimizing temperature for language models with multi-sample inference. *Forty-second international conference on machine learning*. <https://openreview.net/forum?id=rmWpE3FrHW>
- Du, W., & Black, A. W. (2019). Boosting dialog response generation. *57th Annual Meeting of the Association for Computational Linguistics*, 38–43. <https://doi.org/10.18653/v1/P19-1005>
- Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017a). CAN: Creative adversarial networks, generating “art” by learning about styles and deviating from style

- norms. *Proceedings of the 8th International Conference on Computational Creativity*, 96–103.
- Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017b). Can: Creative adversarial networks, generating “art” by learning about styles and deviating from style norms. *Proceedings of the 8th ICCV*, 96–103.
- Eno, B., & Schmidt, P. (1975). Oblique strategies.
- Ens, J., & Pasquier, P. (2018). CAEMSI: A cross-domain analytic evaluation methodology for style imitation. *ICCC*, 64–71.
- Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 889–898). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1082>
- Feldman, D. H., Csikszentmihalyi, M., & Gardner, H. (1994). *Changing the world: A framework for the study of creativity*. Westport: Praeger Publishers.
- Finlayson, M., Hewitt, J., Koller, A., Swayamdipta, S., & Sabharwal, A. (2024). Closing the curious case of neural text degeneration. *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=dONpC9GL1o>
- Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanit. Soc. Sci. Commun.*, 7(1), 1–9.
- Fletcher, R. (2000). Linear programming. In *Practical methods of optimization* (pp. 150–194). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118723203.ch8>
- Franceschelli, G., & Musolesi, M. (2024a). Creative beam search: Llm-as-a-judge for improving response generation (K. Grace, M. T. Llano, P. Martins & M. M. Hedblom, Eds.), 364–368.
- Franceschelli, G., & Musolesi, M. (2024b). Creativity and machine learning: A survey. *ACM Comput. Surv.*, 56(11). <https://doi.org/10.1145/3664595>

- Franceschelli, G., & Musolesi, M. (2025). Diffsampling: Enhancing diversity and accuracy in neural text generation. <https://arxiv.org/abs/2502.14037>
- Fraser, K., Kiritchenko, S., & Nejadgholi, I. (2023). Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 288–292). Association for Computational Creativity.
- Friedman, D., & Dieng, A. B. (2023). The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=g97OHbQyk1>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, 11(2), 127–138.
- Gaier, A., & Ha, D. (2019). Weight agnostic neural networks. *Advances in Neural Information Processing Systems*, 32.
- Galanter, P. (2012). Computational aesthetic evaluation: Past and future. In *Computers and creativity* (pp. 255–293). https://doi.org/10.1007/978-3-642-31727-9_10
- Gallie, W. B. (1955). Essentially contested concepts. *Proceedings of the Aristotelian Society*, 56, 167–198.
- Gamble, C., Gowlett, J., & Dunbar, R. (2014). *Thinking big: How the evolution of social life shaped the human mind*. Thames & Hudson.
- Ganguli, D., Lovitt, L., Kernion, J., Askill, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., ... Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. <https://arxiv.org/abs/2209.07858>

- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Garcia, M. B. (2024). The paradox of artificial creativity: Challenges and opportunities of generative AI artistry. *Creativity Research Journal*, 0(0), 1–14. <https://doi.org/10.1080/10400419.2024.2354622>
- Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.
- Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.
- Gardner, H. (2011). *Creating minds: An anatomy of creativity seen through the lives of freud, einstein, picasso, stravinsky, eliot, graham, and ghandi*. Civitas Books.
- Garimella, A., Banea, C., Hossain, N., & Mihalcea, R. (2020). “judge me by my size (noun), do you?” YodaLib: A demographic-aware humor generation framework. In D. Scott, N. Bel & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 2814–2825). International Committee on Computational Linguistics. <https://doi.org/10.18653/v1/2020.coling-main.253>
- Gell-Mann, M., & Lloyd, S. (2004). Effective complexity. *Nonextensive entropy*, 387–398.
- Gemeinboeck, P., & Saunders, R. (2019). Exploring social co-presence through movement in human robot encounters. *Proceedings of AISB 2019 Symposium on Movement that Shapes Behaviour*.
- Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., ... Andreev, A. (2024). Gemma 2: Improving open language models at a practical size. <https://arxiv.org/abs/2408.00118>
- Gervás, P. (2009). Computational approaches to storytelling and creativity. *AI Magazine*, 30(3), 49. <https://doi.org/10.1609/aimag.v30i3.2250>

- Gervás, P., Pérez & Sosa, R. (2007). On the fly collaborative story-telling: Revising contributions to match a shared partial story line. In A. Cardoso & G. A. Wiggins (Eds.), *Proceedings of the 4th international joint workshop on computational creativity* (pp. 13–20).
- Góes, F., Sawicki, P., Grzes, M., Volpe, M., & Brown, D. (2023a). Is GPT-4 good enough to evaluate jokes? In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 367–371). Association for Computational Creativity.
- Góes, F., Sawicki, P., Grzes, M., Volpe, M., & Watson, J. (2023b). Pushing gpt’s creativity to its limits: Alternative uses and torrance tests. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 342–346). Association for Computational Creativity.
- Goes, F., Zhou, Z., Sawicki, P., Grzes, M., & Brown, D. G. (2022). Crowd Score: A method for the evaluation of jokes using large language model ai voters as judges. *arXiv preprint arXiv:2212.11214*.
- Góngora, S., Chiruzzo, L., Méndez, G., & Gervás, P. (2024). Payador: A minimalist approach to grounding language models on structured data for interactive storytelling and role-playing games. In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 101–106). Association for Computational Creativity.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [<http://www.deeplearningbook.org>]. MIT Press.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
- Gooding, S., & Grefenstette, E. (2025). Interaction dynamics as a reward signal for llms.
- Grace, K., & Maher, M. L. (2014). What to expect when you’re expecting: The role of unexpectedness in computationally evaluating creativity. In S. Colton, D. Ventura,

- N. Lavrač & M. Cook (Eds.), *Proceedings of the fifth international conference on computational creativity* (pp. 120–128).
- Grace, K., & Maher, M. L. (2015). Specific curiosity as a cause and consequence of transformational creativity. In H. Toivonen, S. Colton, M. Cook & D. Ventura (Eds.), *Proceedings of the sixth international conference on computational creativity (iccc 2015)* (pp. 260–267). Brigham Young University.
- Grace, K., & Maher, M. L. (2016). Surprise-triggered reformulation of design goals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). <https://doi.org/10.1609/aaai.v30i1.9890>
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2), 193–202.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). The llama 3 herd of models. <https://arxiv.org/abs/2407.21783>
- Graves, A. (2013). Generating sequences with recurrent neural networks.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357–364. <https://doi.org/10.1016/j.tics.2010.05.004>
- Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., ... & Hajishirzi, H. (2024). OLMo: Accelerating the science of language models. <https://arxiv.org/abs/2402.00838>
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. <https://arxiv.org/abs/2203.05794>
- Grynbaum, M. M., & Mac, R. (2023). The times sues openai and microsoft over a.i. use of copyrighted work. *The New York Times*. Retrieved August 5, 2026, from

<https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>

Guckelsberger, C., Salge, C., & Colton, S. (2017). Addressing the “why?” in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In A. Goel, A. Jordanous & A. Pease (Eds.), *Proceedings of the 8th international conference on computational creativity* (pp. 128–135). ACC.

Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1), 3–14. <https://doi.org/10.1002/j.2162-6057.1967.tb00002.x>

Guo, Y., Shang, G., Vazirgiannis, M., & Clavel, C. (2024). The curious decline of linguistic diversity: Training language models on synthetic text. In K. Duh, H. Gomez & S. Bethard (Eds.), *Findings of the association for computational linguistics: NAACL 2024* (pp. 3589–3604). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.228>

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating large language models in generating synthetic hci research data: A case study. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3544548.3580688>

Hanna, S. (2005). Where creativity comes from. *Proceedings of Human Interaction 2005*, 45–70.

Hantula, O., & Linkola, S. (2018). Towards goal-aware collaboration in artistic agent societies [ICCC, ICC; Conference date: 07-09-2020 Through 11-09-2020]. In F. Pachet, A. Jordanous & C. León (Eds.), *Proceedings of the ninth iccc* (pp. 136–143).

Haque, M. A., & Li, S. (2025). Exploring ChatGPT and its impact on society. *AI and Ethics*, 5(2), 791–803. <https://doi.org/10.1007/s43681-024-00435-4>

Harrison, B., Purdy, C., & Riedl, M. (2021). Toward automated story generation with markov chain monte carlo methods and deep neural networks. *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(2), 191–197. <https://doi.org/10.1609/aiide.v13i2.13003>

- Harrison, J. (2017). *Marjorie prime*. Theatre Communications Group.
- Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2), 100–107. <https://doi.org/10.1109/TSSC.1968.300136>
- Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing.
- Hedderich, M. A., Wang, A., Zhao, R., Eichin, F., Fischer, J., & Plank, B. (2025). What's the difference? supporting users in identifying the effects of prompt and model changes through token patterns. In W. Che, J. Nabende, E. Shutova & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 20093–20123). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.985>
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.
- Heuser, R. (2025). Cultural collapse: Toward a generative formalism for AI cultural production. In T. Arnold, M. Fantoli & R. Ros (Eds.), *Computational humanities research 2025* (pp. 574–587). Anthology of Computers and the Humanities. <https://doi.org/10.63744/USvuyzSlapvy>
- Hewitt, J., Manning, C., & Liang, P. (2022). Truncation sampling as language model desmoothing. In Y. Goldberg, Z. Kozareva & Y. Zhang (Eds.), *Findings of the association for computational linguistics: Emnlp 2022* (pp. 3414–3427). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-emnlp.249>
- Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4129–4138).

- Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1419>
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*. <http://arxiv.org/abs/1503.02531>
- Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02), 107–116.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. *International Conference on Learning Representations*. <https://openreview.net/forum?id=rygGQyrFvH>
- Hughes, E., Dennis, M. D., Parker-Holder, J., Behbahani, F., Mavalankar, A., Shi, Y., Schaul, T., & Rocktäschel, T. (2024). Position: Open-endedness is essential for artificial superhuman intelligence. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett & F. Berkenkamp (Eds.), *Proceedings of the 41st international conference on machine learning* (pp. 20597–20616, Vol. 235). PMLR. <https://proceedings.mlr.press/v235/hughes24a.html>
- Iba, T. (2010). An autopoietic systems theory for creativity. *Procedia - Social and Behavioral Sciences*, 2(4), 6610–6625. <https://doi.org/10.1016/j.sbspro.2010.04.071>
- Ibarrola, F., & Grace, K. (2024). Measuring diversity in co-creative image generation. In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 254–261). Association for Computational Creativity.
- Ibarrola, F., Lawton, T., & Grace, K. (2024). A collaborative, interactive and context-aware drawing agent for co-creative design. *IEEE Transactions on Visualization*

- and Computer Graphics*, 30(8), 5525–5537. <https://doi.org/10.1109/TVCG.2023.3293853>
- IJCAI. (2024). Call for papers – AI, arts & creativity. *IJCAI*. Retrieved May 31, 2025, from <https://ijcai24.org/call-for-papers-ai-arts-creativity>
- Insel, T. R., & Fernald, R. D. (2004). How the brain processes social information: Searching for the social brain. *Annu. Rev. Neurosci.*, 27, 697–722.
- Jacob, A. P., Shen, Y., Farina, G., & Andreas, J. (2024). The consensus game: Language model generation via equilibrium search. *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=n9xeGcI4Yg>
- Jang, M., & Lukasiewicz, T. (2023). Consistency analysis of ChatGPT. In H. Bouamor, J. Pino & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 15970–15985). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.991>
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(s1), 87–106. <https://doi.org/https://doi.org/10.1111/j.1467-9922.2012.00739.x>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023a). Mistral 7b. <https://arxiv.org/abs/2310.06825>
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., ... Sayed, W. E. (2024). Mixtral of experts.
- Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2023b). Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36, 10622–10643.

- Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2023c). Evaluating and inducing personality in pre-trained language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Eds.), *Advances in neural information processing systems* (pp. 10622–10643, Vol. 36). Curran Associates, Inc.
- Johnson, C. G. (2014). Is it time for computational creativity to grow up, start being irresponsible? In S. Colton, D. Ventura, N. Lavrač & M. Cook (Eds.), *Proceedings of the fifth international conference on computational creativity* (pp. 263–267).
- Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computing*, 4(3), 246–279. <https://doi.org/10.1007/s12559-012-9156-1>
- Jordanous, A. (2014, April). What is computational creativity? Retrieved July 3, 2025, from https://www.creativitypost.com/science/what_is_computational_creativity
- Jordanous, A. (2016). Four PPPerspectives on computational creativity in theory and in practice. *Connection Science*, 28(2), 194–216.
- Jordanous, A. (2017). How can we write software that can communicate creatively? *Proceedings of the 4th AISB Symposium on Computational Creativity*.
- Jordanous, A., & Keller, B. (2016). Modelling creativity: Identifying key components through a corpus-based approach. *PloS one*, 11(10), e0162959.
- Jr., A. L., Chakrabarty, T., Hope, T., Pronchick, J., Bhawsar, K., & Beaty, R. E. (2025). How do humans and language models reason about creativity? a comparative analysis. <https://arxiv.org/abs/2502.03253>
- Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd). Online manuscript released January 12, 2025. <https://web.stanford.edu/~jurafsky/slp3/>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

- Kaila, A.-K., Holzapfel, A., & Jääskeläinen, P. (2024). Gardening frictions in creative ai: Emerging art practices and their design implications. In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 186–195). Association for Computational Creativity.
- Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. <https://arxiv.org/abs/2509.04664>
- Kambhampati, S., Valmeekam, K., Bhambri, S., Palod, V., Saldyt, L. P., Stechly, K., Samineni, S. R., Kalwar, D., & Biswas, U. (2026). Position: Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *Forty-third International Conference on Machine Learning Position Paper Track*. <https://openreview.net/forum?id=nP7rL36vYj>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. <https://arxiv.org/abs/2001.08361>
- Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. *Review of General Psychology*, 13(1), 1–12. <https://doi.org/10.1037/a0013688>
- Kaufman, J. C., & Plucker, J. A. (2011). Intelligence and creativity. In R. J. Sternberg & S. B. Kaufman (Eds.), *The cambridge handbook of intelligence* (pp. 771–783). Cambridge University Press.
- Kelly, J., Calderwood, A., Wardrip-Fruin, N., & Mateas, M. (2023). There and back again: Extracting formal domains for controllable neurosymbolic story authoring. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 19(1), 64–74. <https://doi.org/10.1609/aiide.v19i1.27502>
- Khatun, A., & Brown, D. (2023). Reliability check: An analysis of GPT-3's response to sensitive topics and prompt wording. In A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar & R. Gupta (Eds.), *Proceedings of the 3rd workshop on trustworthy natural language processing*

- (TrustNLP 2023) (pp. 73–95). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.trustnlp-1.8>
- Khatun, A., & Brown, D. G. (2024). A study on large language models' limitations in multiple-choice question answering. <https://arxiv.org/abs/2401.07955>
- Kim, J. S., Aheimer, B., Manrara, V. M., & Bedny, M. (2021). Shared understanding of color among sighted and blind adults. *Proceedings of the National Academy of Sciences*, 118(33), e2020192118. <https://doi.org/10.1073/pnas.2020192118>
- Kim, P. J., Jang, Y., Kim, J., & Yoo, J. (2023). TopP&R: Robust support estimation approach for evaluating fidelity and diversity in generative models. *Advances in Neural Information Processing Systems*, 36, 7831–7866.
- Kim, Y., Denton, C., Hoang, L., & Rush, A. M. (2017). Structured attention networks. *International Conference on Learning Representations*. <https://openreview.net/forum?id=HkE0Nvqlg>
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *International Conference on Learning Representations*.
- Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2024). Understanding the effects of RLHF on LLM generalisation and diversity. *International Conference on Learning Representations*. <https://openreview.net/forum?id=PXD3FAVHJT>
- Klein, S., Aeschliman, J. F., Balsiger, D., Converse, S. L., Court, C., Foster, M., Lao, R., Oakley, J. D., & Smith, J. (1973). *Automatic novel writing: A status report* (186). Computer Science Department, The University of Wisconsin. Madison, Wisconsin.
- Knill, D. C., & Pouget, A. (2004). The bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12), 712–719.
- Kojima, T., Gu, S. (, Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho

- & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 22199–22213, Vol. 35). Curran Associates, Inc.
- Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of information transmission*, 1(1), 1–7.
- Koopman, B., & Zuccon, G. (2023). Dr chatgpt tell me what i want to hear: How different prompts impact health answer correctness. In H. Bouamor, J. Pino & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 15012–15022). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.928>
- Kouwenhoven, T. (2025, October). *Collaborative meaning-making: The emergence of novel languages in humans, machines, and human-machine interactions* [Doctoral dissertation, Leiden University] [SIKS Dissertation Series 2025-47]. <https://hdl.handle.net/1887/4281976>
- Kouwenhoven, T., Peeperkorn, M., de Kleijn, R., & Verhoef, T. (2025). Shaping shared languages: Human and large language models' inductive biases in emergent communication. In J. Kwok (Ed.), *Proceedings of the thirty-fourth international joint conference on artificial intelligence, IJCAI-25* (pp. 10298–10306). International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2025/1144>
- Kouwenhoven, T., Peeperkorn, M., Van Dijk, B., & Verhoef, T. (2024). The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke & Y. Oseki (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 57–71). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.cmcl-1.5>
- Kouwenhoven, T., Peeperkorn, M., & Verhoef, T. (2025). Searching for structure: Investigating emergent communication with large language models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 9977–9991). As-

- sociation for Computational Linguistics. <https://aclanthology.org/2025.coling-main.667/>
- Kozbelt, A., Beghetto, R. A., & Runco, M. A. (2010). Theories of creativity. In J. C. Kaufman & R. J. Sternberg (Eds.), *The cambridge handbook of creativity* (pp. 20–47). Cambridge University Press. <https://doi.org/10.1017/CBO9780511763205.004>
- Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.
- Lahikainen, J., Ady, N. M., & Guckelsberger, C. (2024). Creativity and markov decision processes. In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 299–308). Association for Computational Creativity.
- Lai, Y.-A., Zhu, X., Zhang, Y., & Diab, M. (2020). Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. *12th Language Resources and Evaluation Conference*, 1739–1746. <https://aclanthology.org/2020.lrec-1.215>
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- Lamb, C., Brown, D., & Clarke, C. (2016). Evaluating digital poetry: Insights from the CAT. In F. Pachet, A. Cardoso, V. Corruble & F. Ghedini (Eds.), *Proceedings of the seventh international conference on computational creativity* (pp. 60–67). Sony CSL.
- Lamb, C., Brown, D. G., & Clarke, C. (2015). Human competence in creativity evaluation. In H. Toivonen, S. Colton, M. Cook & D. Ventura (Eds.), *Proceedings of the sixth international conference on computational creativity* (pp. 102–109). Brigham Young University.
- Lamb, C., Brown, D. G., & Clarke, C. L. A. (2018). Evaluating computational creativity: An interdisciplinary tutorial. *ACM Comput. Surv.*, 51(2). <https://doi.org/10.1145/3167476>

- Lamb, C., Brown, D. G., & Grossman, M. R. (2025). Precarity and solidarity: Preliminary results on a study of queer and disabled fiction writers' experiences with generative AI. In J. Kwok (Ed.), *Proceedings of the thirty-fourth international joint conference on artificial intelligence, IJCAI-25* (pp. 10117–10125). International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2025/1124>
- Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Bras, R. L., Tafjord, O., Wilhelm, C., Soldaini, L., ... Hajishirzi, H. (2025). Tulu 3: Pushing frontiers in open language model post-training. <https://arxiv.org/abs/2411.15124>
- Latour, B. (1993). *The pasteurization of france*. Harvard University Press.
- Le Bronnec, F., Verine, A., Negrevergne, B., Chevaleyre, Y., & Allauzen, A. (2024). Exploring precision and recall to assess the quality and diversity of LLMs. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11418–11441. <https://doi.org/10.18653/v1/2024.acl-long.616>
- Leinster, T. (2021). *Entropy and diversity: The axiomatic approach*. Cambridge University Press.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 9459–9474, Vol. 33). Curran Associates, Inc.
- Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. <https://doi.org/10.18653/v1/N16-1014>

- Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. M. (2004). The similarity metric. *IEEE transactions on Information Theory*, 50(12), 3250–3264.
- Li, M., Vitányi, P., et al. (2008). *An introduction to kolmogorov complexity and its applications* (Vol. 3). Springer.
- Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., & Lewis, M. (2023). Contrastive decoding: Open-ended text generation as optimization. In A. Rogers, J. Boyd-Graber & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 12286–12312). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.687>
- Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.-J., Chen, X., et al. (2025). From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.
- Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. *Proceedings of the 5th International Conference on Conversational User Interfaces*. <https://doi.org/10.1145/3571884.3604316>
- Lin, Y., Seto, S., ter Hoeve, M., Metcalf, K., Theobald, B.-J., Wang, X., Zhang, Y., Huang, C., & Zhang, T. (2024). On the limited generalization capability of the implicit reward model induced by direct preference optimization. *EMNLP*. <https://arxiv.org/abs/2409.03650>
- Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., ... Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>

- Linkola, S., & Hantula, O. (2018a). On collaborator selection in creative agent societies: An evolutionary art case study. *Computational Intelligence in Music, Sound, Art and Design*, 206–222. https://doi.org/10.1007/978-3-319-77583-8_14
- Linkola, S., & Hantula, O. (2018b). On collaborator selection in creative agent societies: An evolutionary art case study. *Computational Intelligence in Music, Sound, Art and Design*, 206–222. https://doi.org/10.1007/978-3-319-77583-8_14
- Linkola, S., & Hauhio, I. (2023). Many meanings of intentionality: A brief disambiguation for computational creativity. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 153–156). Association for Computational Creativity.
- Linkola, S., & Kantosalo, A. (2019). Extending the creative systems framework for the analysis of creative agent societies. *Proceedings of the 10th ICCG*.
- Liu, S., Sabour, S., Zheng, Y., Ke, P., Zhu, X., & Huang, M. (2022). Rethinking and refining the distinct metric. *60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 762–770. <https://doi.org/10.18653/v1/2022.acl-short.86>
- Luhmann, N. (1984). Soziale systeme: Grundriss einer allgemeinen theorie.
- Lyu, S., Rockmore, D., & Farid, H. (2004). A digital technique for art authentication. *Proceedings of the National Academy of Sciences*, 101(49), 17006–17010.
- Ma, Q. C., Wu, S. T., & Koedinger, K. (2023). Is AI the better programming partner? human-human pair programming vs. human-AI pair programming. In S. Moore, J. Stamper, R. Tong, C. Cao, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh & C. Brooks (Eds.), *Proceedings of the workshop on empowering education with llms – the next-gen interface and content generation 2023* (Vol. 3487).
- Machado, P., Romero, J., & Manaris, B. (2008). Experiments in computational aesthetics. In J. Romero & P. Machado (Eds.), *The art of artificial evolution: A handbook on evolutionary art and music* (pp. 381–415). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72877-1_18

- Manaris, B., Vaughan, D., Wagner, C., Romero, J., & Davis, R. B. (2003). Evolutionary music and the zipf-mandelbrot law: Developing fitness functions for pleasant music. In S. Cagnoni, C. G. Johnson, J. J. R. Cardalda, E. Marchiori, D. W. Corne, J.-A. Meyer, J. Gottlieb, M. Middendorf, A. Guillot, G. R. Raidl & E. Hart (Eds.), *Applications of evolutionary computing* (pp. 522–534). Springer Berlin Heidelberg.
- Manjavacas, E., Karsdorp, F., Burtenshaw, B., & Kestemont, M. (2017). Synthetic literature: Writing science fiction in a co-creative process. *Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, 29–37. <https://doi.org/10.18653/v1/W17-3904>
- Manning, C. D., & Schütze, H. (1999, May). *Foundations of statistical natural language processing*. MIT Press.
- Manta, A. (2018). Demystifying creativity: An assemblage perspective towards artistic creativity. *Creativity Studies*, 11(1), 85–101. <https://doi.org/10.3846/cs.2018.542>
- Marcus, G. (2025). The most underreported and important story in AI right now is that pure scaling has failed to produce AGI. Retrieved May 31, 2025, from <https://fortune.com/2025/02/19/generative-ai-scaling-agi-deep-learning/>
- Martindale, C. (1990a). *The clockwork muse: The predictability of artistic change*. Basic Books.
- Martindale, C. (1990b). *The clockwork muse: The predictability of artistic change*. Basic Books.
- Maturana, H. R., & Varela, F. J. (1980). Autopoiesis and cognition: The realization of living. *Boston Studies in the Philosophy of Science*, 42.
- McCormack, J., & Cruz Gambardella, C. (2022). Complexity and aesthetics in generative and evolutionary art. *Genetic Programming and Evolvable Machines*. <https://doi.org/10.1007/s10710-022-09429-9>
- McCormack, J., Cruz Gambardella, C., Rajcic, N., Krol, S. J., Llano, M. T., & Yang, M. (2023). Is writing prompts really making art? In C. Johnson, N. Rodríguez-Fernández & S. M. Rebelo (Eds.), *Artificial intelligence in music, sound, art and design* (pp. 196–211). Springer Nature Switzerland.

- McCormack, J., Wilson, E., Rajcic, N., & Llano, M. T. (2024). Mimetic poet. In K. Grace, M. T. Llano, P. Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 279–288). Association for Computational Creativity.
- McGraw, G., & Hofstadter, D. (1993). Letter spirit: An architecture for creativity in a microdomain. In P. Torasso (Ed.), *Advances in artificial intelligence* (pp. 65–70). Springer Berlin Heidelberg.
- Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(3), 333.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, 85(3), 207.
- Meehan, J. R. (1977a). TALE-SPIN, an interactive program that writes stories. *5th International Joint Conference on Artificial Intelligence*, 91–98.
- Meehan, J. R. (1977b). TALE-SPIN, an interactive program that writes stories. *5th International Joint Conference on Artificial Intelligence*, 91–98.
- Meister, C., Pimentel, T., Wiher, G., & Cotterell, R. (2023). Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11, 102–121. https://doi.org/10.1162/tacl_a_00536
- Meister, C., Vieira, T., & Cotterell, R. (2020). Best-first beam search (M. Johnson, B. Roark & A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, 8, 795–809. https://doi.org/10.1162/tacl_a_00346
- Méndez, G., & Gervás, P. (2023). Using ChatGPT for story sifting in narrative generation. *14th International Conference on Computational Creativity*, 387–391.
- Miaschi, A., Brunato, D., Dell’Orletta, F., & Venturi, G. (2021). What makes my model perplexed? a linguistic investigation on neural language models perplexity. In E. Agirre, M. Apidianaki & I. Vulić (Eds.), *Proceedings of deep learning inside out*

- (*deelio*): *The 2nd workshop on knowledge extraction and integration for deep learning architectures* (pp. 40–47). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.deelio-1.5>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. <https://arxiv.org/abs/1301.3781>
- Millière, R. (2024). Language models as models of language. <https://arxiv.org/abs/2408.07144>
- Mollo, D. C., & Millière, R. (2023). The vector grounding problem. <https://arxiv.org/abs/2304.01481>
- Mondol, T., & Brown, D. (2021a). Incorporating algorithmic information theory into fundamental concepts of computational creativity. In A. G. de Silva Garza, T. Veale, W. Aguilar & R. P. y Pérez (Eds.), *Proceedings of the 12th international conference on computational creativity* (pp. 173–181). Association for Computational Creativity.
- Mondol, T., & Brown, D. G. (2021b). Computational creativity and aesthetics with algorithmic information theory. *Entropy*, 23(12), 1654, 1–24. <https://doi.org/10.3390/e23121654>
- Moradi Dakhel, A., Majdinasab, V., Nikanjam, A., Khomh, F., Desmarais, M. C., & Jiang, Z. M. (2023). GitHub Copilot AI pair programmer: Asset or liability? *Journal of Systems and Software*, 203, 111734. <https://doi.org/10.1016/j.jss.2023.111734>
- Morain, R., Kinghorn, B., & Ventura, D. (2023). Are language models unsupervised multi-domain CC systems? In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 39–43). Association for Computational Creativity.
- Morin, O., Jacquet, P. O., Vaesen, K., & Acerbi, A. (2021). Social information use and social information waste. *Phil. Trans. R. Soc. B*, 376(1828), 20200052.
- Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024). Position: Levels of agi for operationalizing progress on the path

- to agi. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett & F. Berkenkamp (Eds.), *Proceedings of the 41st international conference on machine learning* (pp. 36308–36321, Vol. 235). PMLR. <https://proceedings.mlr.press/v235/morris24b.html>
- Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). Mteb: Massive text embedding benchmark. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2014–2037). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.eacl-main.148>
- Nadeem, M., He, T., Cho, K., & Glass, J. (2020). A systematic characterization of sampling algorithms for open-ended language generation. *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 334–346. <https://aclanthology.org/2020.aacl-main.36>
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), 133–142. <https://doi.org/10.1111/j.2041-210x.2012.00261.x>
- NeurIPS. (2023). NeurIPS creative AI track: Call for proposal – theme: Celebrating diversity. *NeurIPS*. Retrieved May 31, 2025, from <https://neurips.cc/Conferences/2023/CallForCreativeAI>
- Novikov, A., Vű, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J. R., Mehrabian, A., Kumar, M. P., See, A., Chaudhuri, S., Holland, G., Davies, A., Nowozin, S., Kohli, P., & Balog, M. (2025, May). *Alphae-volve: A coding agent for scientific and algorithmic discovery* (tech. rep.). Google Deepmind.
- O’Brien, S., & Lewis, M. (2023). Contrastive decoding improves reasoning in large language models [Preprint]. <https://arxiv.org/abs/2309.09117>

- Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25), e2022340118. <https://doi.org/10.1073/pnas.2022340118>
- OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Proceedings of the 36th International Conference on Neural Information Processing Systems*.
- Padmakumar, V., & He, H. (2024). Does writing with language models reduce content diversity? *12th International Conference on Learning Representations*. <https://openreview.net/forum?id=Feiz5HtCD0>
- Pal, A., Karkhanis, D., Dooley, S., Roberts, M., Naidu, S., & White, C. (2024). Smaug: Fixing failure modes of preference optimisation with dpo-positive [Preprint]. <https://arxiv.org/abs/2402.13228>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318. <https://doi.org/10.3115/1073083.1073135>
- Parikh, A., Täckström, O., Das, D., & Uszkoreit, J. (2016). A decomposable attention model for natural language inference. In J. Su, K. Duh & X. Carreras (Eds.), *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2249–2255). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D16-1244>
- Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. <https://doi.org/10.1145/3586183.3606763>

- Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, 1423(1), 378–395. <https://doi.org/https://doi.org/10.1111/nyas.13654>
- Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* [Doctoral dissertation, Department of Computer Science, City University of London].
- Pease, A., & Colton, S. (2011a). Computational creativity theory: Inspirations behind the FACE and the IDEA models. In D. Ventura, P. Gervás, D. F. Harrell, M. L. Maher, A. Pease & G. Wiggins (Eds.), *Proceedings of the second international conference on computational creativity* (pp. 72–77).
- Pease, A., & Colton, S. (2011b). On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. *Proceedings of the AISB symposium on AI and Philosophy*, 39.
- Peeperkorn, M. (2022). Artificial Creative Societies: Adaption, Intention, and Evaluation. *Creativity and Cognition*, 704–707. <https://doi.org/10.1145/3527927.3533728>
- Peeperkorn, M., Bown, O., & Saunders, R. (2020). The maintenance of conceptual spaces through social interactions [Master Thesis Abstract]. *BNAIC/BeneLearn 2020*, 430–431.
- Peeperkorn, M., Brown, D. G., & Jordanous, A. (2023). On characterizations of large language models and creativity evaluation. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 143–147). Association for Computational Creativity.
- Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2025). Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models [Preprint, Under review]. <https://arxiv.org/abs/2507.20956>
- Peeperkorn, M., Kouwenhoven, T., Brown, D. G., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models? In K. Grace, M. T. Llano, P.

- Martins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 226–235). Association for Computational Creativity.
- Peeperkorn, M., Saunders, R., Bown, O., & Jordanous, A. (2022). Mechanising conceptual spaces using variational autoencoders. In M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz & T. Veale (Eds.), *Proceedings of the 13th international conference on computational creativity* (pp. 287–290). Association for Computational Creativity.
- Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. In Y. Goldberg, Z. Kozareva & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 3419–3448). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.225>
- Pérez y Pérez, R., & Sharples, M. (2001). MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(2), 119–139. <https://doi.org/10.1080/09528130010029820>
- Perrigo, B. (2023, January). Exclusive: OpenAI used kenyan workers on less than \$2 per hour to make ChatGPT less toxic. Retrieved May 24, 2023, from <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Phuong, M., & Hutter, M. (2022). Formal algorithms for transformers.
- Pillutla, K., Liu, L., Thickstun, J., Welleck, S., Swayamdipta, S., Zellers, R., Oh, S., Choi, Y., & Harchaoui, Z. (2023). Mauve scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356), 1–92. <http://jmlr.org/papers/v24/23-0023.html>
- Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., & Harchaoui, Z. (2021). MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (pp. 4816–4828, Vol. 34). Curran Associates, Inc.

- Pimentel, T., Meister, C. I., & Cotterell, R. (2023). On the usefulness of embeddings, clusters and strings for text generation evaluation. *International Conference on Learning Representations*. <https://openreview.net/forum?id=bvpkw7UIRdU>
- Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. University of Chicago Press.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.*, 1(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- 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., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 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
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *37th International Conference on Neural Information Processing Systems*.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. <https://arxiv.org/abs/2204.06125>
- Ravfogel, S., Goldberg, Y., & Goldberger, J. (2023). Conformal nucleus sampling. *Findings of the Association for Computational Linguistics: ACL 2023*, 27–34. <https://doi.org/10.18653/v1/2023.findings-acl.3>
- Ray, T. S. (1991). Evolution and optimization of digital organisms. *Scientific excellence in supercomputing: The IBM 1990 contest prize papers*.
- Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, 44(3), 393–401. https://doi.org/10.1162/coli_a_00322

- Ren, Y., Jin, R., Zhang, T., & Xiong, D. (2025). Do large language models mirror cognitive language processing? In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 2988–3001). Association for Computational Linguistics. <https://aclanthology.org/2025.coling-main.201/>
- Renze, M. (2024). The effect of sampling temperature on problem solving in large language models. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 7346–7356). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.432>
- Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan*, 42(7), 305–310.
- Riedl, M. O., & Young, R. M. (2006). Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing*, 24(3), 303–323. <https://doi.org/10.1007/BF03037337>
- Rigau, J., Feixas, M., & Sbert, M. (2007). Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. *Proceedings of the Third Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 105–112.
- Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds & Machines*, 17, 76–99.
- Robinson, K., & Brown, D. G. (2021). Quantitative user perceptions of music recommendation list diversity. *ISMIR*, 562–568.
- Robinson, K., Brown, D. G., & Schedl, M. (2020). User insights on diversity in music recommendation lists. *ISMIR*, 446–453.
- Roemmele, M., & Gordon, A. S. (2018). Automated assistance for creative writing with an rnn language model. *23rd International Conference on Intelligent User Interfaces Companion*. <https://doi.org/10.1145/3180308.3180329>

- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685. <https://doi.org/10.1109/CVPR52688.2022.01042>
- Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., & Fawzi, A. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625(7995), 468–475. <https://doi.org/10.1038/s41586-023-06924-6>
- Romero, J., Machado, P., Carballal, A., & Correia, J. (2012). Computing aesthetics with image judgement systems. In J. McCormack & M. d’Inverno (Eds.), *Computers and creativity* (pp. 295–322). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31727-9_11
- Roose, K. (2025). For some recent graduates, the A.I. job apocalypse may already be here. Retrieved May 31, 2025, from <https://www.nytimes.com/2025/05/30/technology/ai-jobs-college-graduates.html>
- Rosch, E. (1973). Natural categories. *Cognitive Psychology*, 4(3), 328–350.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382–439.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>
- Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92–96. <https://doi.org/10.1080/10400419.2012.650092>
- Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 31.

- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., & Chen, X. (2016). Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.
- Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6(5), 309–316.
- Saunders, R. (2011). Artificial creative systems and the evolution of language. *Proceedings of the Second ICCS*, 36–41.
- Saunders, R. (2012). Towards autonomous creative systems: A computational approach. *Cognitive Computation*, 4(3), 216–225.
- Saunders, R., & Bown, O. (2015). Computational social creativity. *Artificial Life*, 21(3), 366–378. https://doi.org/10.1162/ARTL_a_00177
- Saunders, R., & Gero, J. S. (2001a). The digital clockwork muse: A computational model of aesthetic evolution. *Proceedings of the AISB'01 Symposium on AI and Creativity in Arts and Science*, 1, 12–21.
- Saunders, R., & Gero, J. S. (2001b). The digital clockwork muse: A computational model of aesthetic evolution. *Proceedings of the AISB*, 1, 12–21.
- Saunders, R., & Grace, K. (2008). Towards a computational model of creative cultures. *2008 AAAI Spring Symposium: Creative Intelligent Systems*.
- Sawicki, P., Grześ, M., Brown, D. G., & Góes, F. (2025). Can large language models outperform non-experts in poetry evaluation? a comparative study using the consensual assessment technique. <https://arxiv.org/abs/2502.19064>
- Sawicki, P., Grześ, M., Goes, F., Brown, D. G., Peeperkorn, M., & Khatun, A. (2023a). Bits of grass: Does GPT already know how to write like whitman? In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 317–321).
- Sawicki, P., Grześ, M., Goes, F., Brown, D. G., Peeperkorn, M., Khatun, A., & Paraskevopoulou, S. (2023b). On the power of special-purpose GPT models to create and

- evaluate new poetry in old styles. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 10–19).
- Sawicki, P., Grześ, M., Jordanous, A., Brown, D. G., & Peeperkorn, M. (2022). Training GPT-2 to represent two romantic-era authors: Challenges, evaluations and pitfalls. In M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz & T. Veale (Eds.), *Proceedings of the 13th international conference on computational creativity* (pp. 34–43).
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3), 230–247. <https://doi.org/10.1109/TAMD.2010.2056368>
- Schmidhuber, J. (2012). A formal theory of creativity to model the creation of art. In J. McCormack & M. d’Inverno (Eds.), *Computers and creativity* (pp. 323–337). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31727-9_12
- Schoenegger, P., Salvi, F., Liu, J., Nan, X., Debnath, R., Fasolo, B., Leivada, E., Recchia, G., Günther, F., Zarifhonarvar, A., Kwon, J., Islam, Z. U., Dehnert, M., Lee, D. Y. H., Reinecke, M. G., Kamper, D. G., Kobaş, M., Sandford, A., Kgomo, J., ... Karger, E. (2026, March). When Large Language Models are More Persuasive Than Incentivized Humans, and Why. <https://doi.org/10.48550/arXiv.2505.09662>
- Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118. <https://doi.org/10.1073/pnas.2105646118>
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms [arXiv:707.06347]. <https://arxiv.org/abs/1707.06347>
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *9th Python in Science Conference*.

- Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Shaib, C., Barrow, J., Sun, J., Siu, A. F., Wallace, B. C., & Nenkova, A. (2025). Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. <https://arxiv.org/abs/2403.00553>
- Shanahan, M. (2024). Talking about large language models. *Commun. ACM*, 67(2), 68–79. <https://doi.org/10.1145/3624724>
- Shao, Y., Li, L., Dai, J., & Qiu, X. (2023). Character-LLM: A trainable agent for role-playing. In H. Bouamor, J. Pino & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 13153–13187). Association for Computational Linguistics. <https://aclanthology.org/2023.emnlp-main.814/>
- Sharples, M., & Pérez y Pérez, R. (2022). *Story machines: How computers have become creative writers* (1st). Routledge. <https://doi.org/10.4324/9781003161431>
- Shazeer, N., Mirhoseini, *, Maziarz, *, Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations*.
- Shihadeh, J., & Ackerman, M. (2023a). Shattering bias: A path to bridging the gender divide with creative machines. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 278–282). Association for Computational Creativity.
- Shihadeh, J., & Ackerman, M. (2023b). What does genius look like? an analysis of brilliance bias in text-to-image models. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 235–244). Association for Computational Creativity.
- Shneiderman, B. (2002). Creativity support tools. *Commun. ACM*, 45(10), 116–120. <https://doi.org/10.1145/570907.570945>

- Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). Ai models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759.
- Shypula, A., Li, S., Zhang, B., Padmakumar, V., Yin, K., & Bastani, O. (2025). Does instruction tuning reduce diversity? a case study using code generation copy. *The 3rd DL4C Workshop: Emergent Possibilities and Challenges in Deep Learning for Code*. <https://openreview.net/forum?id=40uDwtrbd3>
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>
- Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process darwinian? *Psychological Inquiry*, 10(4), 309–328.
- Simonton, D. K. (2012). Taking the U.S. patent office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity Research Journal*, 24(2-3), 97–106. <https://doi.org/10.1080/10400419.2012.676974>
- Simonton, D. K. (2023). The blind-variation and selective-retention theory of creativity: Recent developments and current status of BVSR. *Creativity Research Journal*, 35(3), 304–323. <https://doi.org/10.1080/10400419.2022.2059919>
- Smith, T. C., & Witten, I. H. (1991a). A planning mechanism for generating story text. *Literary and Linguistic Computing*, 6(2), 119–126.
- Smith, T. C., & Witten, I. H. (1991b). A planning mechanism for generating story text. *Literary and Linguistic Computing*, 6(2), 119–126.

- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 2256–2265.
- Sood, A., Llano, M. T., & McCormack, J. (2025). Do conversational interfaces limit creativity? exploring visual graph systems for creative writing. In *Proceedings of the 16th international conference on computational creativity*. Association for Computational Creativity.
- Standish, R. K. (2003). Open-ended artificial evolution. *International Journal of Computational Intelligence and Applications*, 03(02), 167–175. <https://doi.org/10.1142/S1469026803000914>
- Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary computation*, 10(2), 99–127.
- Stasaski, K., & Hearst, M. (2022). Semantic diversity in dialogue with natural language inference. In M. Carpuat, M.-C. de Marneffe & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 85–98). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.6>
- Stein, M. I. (1953). Creativity and culture. *The Journal of Psychology*, 36(2), 311–322. <https://doi.org/10.1080/00223980.1953.9712897>
- Sternberg, R. J., & O'Hara, L. A. (1998). Creativity and intelligence. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 251–272). Cambridge University Press. <https://doi.org/10.1017/CBO9780511807916.015>
- Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). Putting GPT-3's creativity to the (alternative uses) test. In M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz & T. Veale (Eds.), *Proceedings of the 13th international conference on computational creativity* (pp. 164–168). Association for Computational Creativity.

- Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a “horse”. *IEEE Transactions on Multimedia*, 16(6), 1636–1644. <https://doi.org/10.1109/TMM.2014.2330697>
- Sturua, S., Mohr, I., Akram, M. K., Günther, M., Wang, B., Krimmel, M., Wang, F., Mastropas, G., Koukounas, A., Koukounas, A., Wang, N., & Xiao, H. (2024). jina-embeddings-v3: Multilingual embeddings with task LoRA. <https://arxiv.org/abs/2409.10173>
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc.
- Sutton, R. (2019). The bitter lesson. Retrieved April 30, 2025, from <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>
- Taneja, K., Segal, R., & Goodwin, R. (2023). Monte carlo tree search for recipe generation using GPT-2. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 20–28). Association for Computational Creativity.
- Team OLMo, Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., ... Hajishirzi, N. A. S. H. (2025). 2 OLMo 2 furious. <https://arxiv.org/abs/2501.00656>
- Tekin, S. F., İlhan, F., Huang, T., Hu, S., & Liu, L. (2024). LLM-TOPLA: Efficient LLM ensemble by maximising diversity. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 11951–11966. <https://doi.org/10.18653/v1/2024.findings-emnlp.698>
- Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In A. Korhonen, D. Traum & L. Màrquez (Eds.), *Proceedings of the 57th annual*

- meeting of the association for computational linguistics* (pp. 4593–4601). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1452>
- Tevet, G., & Berant, J. (2021). Evaluating the evaluation of diversity in natural language generation. *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 326–346. <https://doi.org/10.18653/v1/2021.eacl-main.25>
- Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science (New York, NY)*, 379(6630), 313. <https://doi.org/10.1126/science.adg7879>
- Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press. <https://doi.org/10.2307/j.ctvjsf4jc>
- Toplyn, J. (2022). Witscript 2: A system for generating improvised jokes without wordplay. In M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz & T. Veale (Eds.), *Proceedings of the 13th international conference on computational creativity* (pp. 54–58). Association for Computational Creativity.
- Torrance, E. P. (1966). *Torrance tests of creative thinking: Norms-technical manual*. Personnel Press.
- Torrance, E. P. (1972). Predictive validity of the torrance tests of creative thinking. *The Journal of Creative Behavior*, 6(4), 236–262. <https://doi.org/https://doi.org/10.1002/j.2162-6057.1972.tb00936.x>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023a). LLaMA: Open and efficient foundation language models. <https://arxiv.org/abs/2302.13971>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023b). Llama 2: Open foundation and fine-tuned chat models. <https://arxiv.org/abs/2307.09288>

- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., & MacDiarmid, M. (2023). Activation addition: Steering language models without optimization [arXiv:2308.10248].
- Turner, S. R. (1994). *The creative process: A computer model of storytelling and creativity*. Lawrence Erlbaum Associates.
- Udandaraao, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P. H. S., Bibi, A., Albanie, S., & Bethge, M. (2024). No “zero-shot” without exponential data: Pretraining concept frequency determines multimodal model performance.
- van Dijk, B., Kouwenhoven, T., Spruit, M., & van Duijn, M. J. (2023). Large language models: The need for nuance in current debates and a pragmatic perspective on understanding. In H. Bouamor, J. Pino & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 12641–12654). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.779>
- van Duijn, M. J. (2016). *The lazy mindreader: A humanities perspective on mindreading and multiple-order intentionality* [Doctoral dissertation, Leiden University]. <https://hdl.handle.net/1887/38817>
- van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., & van der Putten, P. (2023). Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In J. Jiang, D. Reitter & S. Deng (Eds.), *Proceedings of the 27th conference on computational natural language learning (CoNLL)* (pp. 389–402). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.conll-1.25>
- van Stegeren, J., & Theune, M. (2019). Narrative generation in the wild: Methods from NaNoGenMo. *Second Workshop on Storytelling*, 65–74. <https://doi.org/10.18653/v1/W19-3407>

- van Stein, N., & Bäck, T. (2025). LLaMEA: A large language model evolutionary algorithm for automatically generating metaheuristics. *IEEE Transactions on Evolutionary Computation*, 29(2), 331–345. <https://doi.org/10.1109/TEVC.2024.3497793>
- van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *12th International Conference on Natural Language Generation*, 355–368. <https://doi.org/10.18653/v1/W19-8643>
- van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, 614(7947), 224–226.
- VandeHei, J., & Allen, M. (2025). Ai jobs danger: Sleepwalking into a white-collar blood-bath. Retrieved June 3, 2025, from <https://www.axios.com/2025/05/28/ai-jobs-white-collar-unemployment-anthropic>
- Varshney, L. R. (2020). Limits theorems for creativity with intentionality. In T. V. F. Amílcar Cardoso Penousal Machado & J. M. Cunha (Eds.), *Proceedings of the 11th international conference on computational creativity* (pp. 390–393). Association for Computational Creativity.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.
- Veale, T. (2015). Scoffing at mere generation. Retrieved May 7, 2025, from <https://web.archive.org/web/20170402185912/https://prosecco-network.eu/blog/scoffing-mere-generation>
- Veale, T. (2022). Two-fisted comics generation: Comics as a medium and as a representation for creative meanings. In M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz & T. Veale (Eds.), *Proceedings of the 13th international conference on computational creativity* (pp. 59–66). Association for Computational Creativity.
- Veale, T. (2024). From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with llms. In K. Grace, M. T. Llano, P. Mar-

- tins & M. M. Hedblom (Eds.), *Proceedings of the 15th international conference on computational creativity* (pp. 236–244).
- Ventura, D. (2016). Mere generation: Essential barometer or dated concept. In F. Pachet, A. Cardoso, V. Corruble & F. Ghedini (Eds.), *Proceedings of the 7th international conference on computational creativity* (pp. 17–24).
- Ventura, D. (2019). Autonomous intentionality in computationally creative systems. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems* (pp. 49–69). Springer International Publishing. https://doi.org/10.1007/978-3-319-43610-4_3
- Ventura, D. (2023). The emperor's new co-author. In A. Pease, J. M. Cunha, M. Ackerman & D. G. Brown (Eds.), *Proceedings of the 14th international conference on computational creativity* (pp. 55–63).
- Ventura, D. (2026). A Process-centric Analysis of Modern Generative AI As Computational Creativity. In Hugo Goncalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado & Ana Alves (Eds.), *Proceedings of the 17th International Conference on Computational Creativity*. Association for Computational Creativity.
- Vidra, I. D., Kimron, G., Noy, L., & Shamir, A. (2025). Playing along: Building ai agents for co-creation of improvised stories. In *Proceedings of the 16th international conference on computational creativity*. Association for Computational Creativity.
- Vijayakumar, A., Cogswell, M., Selvaraju, R., Sun, Q., Lee, S., Crandall, D., & Batra, D. (2018). Diverse beam search for improved description of complex scenes. *AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.12340>
- Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the Proceedings of the 25th International Conference on Machine Learning*, 1096–1103. <https://doi.org/10.1145/1390156.1390294>

- Vitányi, P. M. (2020). How Incomputable Is Kolmogorov Complexity? *Entropy*, 22(4). <https://doi.org/10.3390/e22040408>
- Volum, R., Rao, S., Xu, M., DesGarennes, G., Brockett, C., Van Durme, B., Deng, O., Malhotra, A., & Dolan, B. (2022). Craft an iron sword: Dynamically generating interactive game characters by prompting large language models tuned on code. *3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)*, 25–43. <https://doi.org/10.18653/v1/2022.wordplay-1.3>
- Vygotsky, L. S. (2004). Imagination and creativity in childhood. *Journal of Russian & East European Psychology*, 42(1), 7–97. <https://doi.org/10.1080/10610405.2004.11059210>
- Walia, C. (2019). A dynamic definition of creativity. *Creativity Research Journal*, 31(3), 237–247. <https://doi.org/10.1080/10400419.2019.1641787>
- Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526–1541. <https://doi.org/10.1038/s41562-023-01659-w>
- Weber, L., Bruni, E., & Hupkes, D. (2023). Mind the instructions: A holistic evaluation of consistency and interactions in prompt-based learning. *27th Conference on Computational Natural Language Learning (CoNLL)*, 294–313. <https://doi.org/10.18653/v1/2023.conll-1.20>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.
- Welleck, S., Bertsch, A., Finlayson, M., Schoelkopf, H., Xie, A., Neubig, G., Kulikov, I., & Harchaoui, Z. (2024). From decoding to meta-generation: Inference-time

- algorithms for large language models. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=eskQMclbMS>
- Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2020). Neural text generation with unlikelihood training. *International Conference on Learning Representations*. <https://openreview.net/forum?id=SJeYe0NtvH>
- Wiggins, G. A. (2006a). A preliminary framework for description, analysis, and comparison of creative systems. *Knowledge-Based Systems*, 19(7), 449–458. <https://doi.org/https://doi.org/10.1016/j.knosys.2006.04.009>
- Wiggins, G. A. (2006b). Searching for computational creativity. *New Generation Computing*, 24(3), 209–222.
- Wiggins, G. A. (2019). A framework for description, analysis, and comparison of creative systems. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems* (pp. 21–47). Springer International Publishing. https://doi.org/10.1007/978-3-319-43610-4_2
- Wiher, G., Meister, C., & Cotterell, R. (2022). On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10, 997–1012. https://doi.org/10.1162/tacl_a_00502
- Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.
- Xin, H., Ren, Z. Z., Song, J., Shao, Z., Zhao, W., Wang, H., Liu, B., Zhang, L., Lu, X., Du, Q., Gao, W., Zhu, Q., Yang, D., Gou, Z., Wu, Z. F., Luo, F., & Ruan, C. (2024). DeepSeek-Prover-V1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. <https://arxiv.org/abs/2408.08152>
- Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Durme, B. V., Murray, K., & Kim, Y. J. (2024a). Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Forty-first international conference on machine learning*. <https://openreview.net/forum?id=51iwkioZpn>

- Xu, W., Jojic, N., Rao, S., Brockett, C., & Dolan, B. (2024b). Echoes in AI: Quantifying lack of plot diversity in LLM outputs. <https://arxiv.org/abs/2501.00273>
- Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., Yan, B., & Chen, Y. (2024). Survey on knowledge distillation for large language models: Methods, evaluation, and application. *ACM Trans. Intell. Syst. Technol.* <https://doi.org/10.1145/3699518>
- Yang, Z., Dai, Z., Salakhutdinov, R., & Cohen, W. W. (2018). Breaking the softmax bottleneck: A high-rank RNN language model. *International Conference on Learning Representations*. <https://openreview.net/forum?id=HkwZSG-CZ>
- Ye, A., Cui, C. Z., Shi, T., & Riedl, M. (2023). Neural story planning. *The AAAI-23 Workshop on Creative AI Across Modalities*. <https://openreview.net/forum?id=cLBKlu5WZK>
- Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story writing with large language models. *27th International Conference on Intelligent User Interfaces*, 841–852. <https://doi.org/10.1145/3490099.3511105>
- Zenil, H., Delahaye, J.-P., & Gaucherel, C. (2012). Image characterization and classification by physical complexity. *Complexity*, 17(3), 26–42. <https://doi.org/https://doi.org/10.1002/cplx.20388>
- Zhang, C., Cheng, J., Shumailov, I., Constantinides, G., & Zhao, Y. (2023). Revisiting block-based quantisation: What is important for sub-8-bit LLM inference? *2023 Conference on Empirical Methods in Natural Language Processing*, 9988–10006. <https://doi.org/10.18653/v1/2023.emnlp-main.617>
- Zhang, H., Duckworth, D., Ippolito, D., & Neelakantan, A. (2021). Trading off diversity and quality in natural language generation. *Workshop on Human Evaluation of NLP Systems (HumEval)*, 25–33. <https://aclanthology.org/2021.humeval-1.3>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson,

- K. Saenko, M. Hardt & S. Levine (Eds.), *Advances in neural information processing systems* (pp. 46595–46623, Vol. 36). Curran Associates, Inc.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., YU, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023a). LIMA: Less is more for alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Eds.), *Advances in neural information processing systems* (pp. 55006–55021, Vol. 36). Curran Associates, Inc.
- Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, 3(3). <https://doi.org/10.1093/pnasnexus/pgae052>
- Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023b). Instruction-following evaluation for large language models. <https://arxiv.org/abs/2311.07911>
- Zhu, W., Hao, H., He, Z., Ai, Y., & Wang, R. (2024). Improving open-ended text generation via adaptive decoding. *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=aXD94eATtT>
- Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., & Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. *The 41st international ACM SIGIR conference on research & development in information retrieval*, 1097–1100. <https://doi.org/10.1145/3209978.3210080>
- Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., & Shen, C. (2023). A survey on efficient training of transformers. *Proceedings of the thirty-second international joint conference on artificial intelligence*. <https://doi.org/10.24963/ijcai.2023/764>
- Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., ... Hendrycks, D. (2023). Representation engineering: A top-down approach to AI transparency [arXiv:2310.01405].

Appendix A

Supplementary Materials for Chapter 3

A.1 The Spectrum of the Mere Generation Prototypical Algorithms

Algorithm 1 Generation using a stochastic process.

```
1: create( $I$ )
2:    $a \leftarrow \{\}$ 
3:   while not done do
4:      $a \leftarrow a + \text{random\_atom}()$ 
5:   return  $a$ 
```

Algorithm 2 Generation by plagiarising an inspiring set I .

```
1: create( $I$ )
2:    $a \leftarrow \text{random\_select}(I)$ 
3:   return  $a$ 
```

Algorithm 3 Generation by memorising an inspiring set I .

```

1: create( $I$ )
2:   $model \leftarrow \text{build\_model}(I)$ 
3:   $a \leftarrow \text{random\_from\_memory}(model)$ 
4:  return  $a$ 

```

Algorithm 4 Generation by generalising an inspiring set I .

```

1: create( $I$ )
2:   $model \leftarrow \text{build\_model}(I)$ 
3:   $a \leftarrow \text{generalize\_from\_model}(model)$ 
4:  return  $a$ 

```

Algorithm 5 Generation by modelling an inspiring set I and filtering candidate artefacts via a fitness function $\text{fitness}()$.

```

1: create( $I$ ,  $\text{fitness}()$ )
2:   $model \leftarrow \text{build\_model}(I)$ 
3:  while  $score < \theta$  do
4:     $a = \text{generate}(model)$ 
5:     $score = \text{fitness}(a)$ 
6:  return  $a$ 

```

Algorithm 6 Knowledge-based generation by modelling an inspiring set I , employing a fitness function $\text{fitness}()$ and leveraging a knowledge base K .

```

1: create( $I$ ,  $\text{fitness}()$ ,  $K$ )
2:   $model \leftarrow \text{build\_model}(I)$ 
3:  while  $score < \theta$  do
4:     $a = \text{generate}(model, K)$ 
5:     $score = \text{fitness}(a)$ 
6:  return  $a$ 

```

Algorithm 7 Creative generation by modelling an inspiring set I and leveraging a knowledge base K followed by evaluating the perception of the generated artefact.

```

1: create( $I, K$ )
2:    $model \leftarrow \text{build\_model}(I)$ 
3:   while  $score < \theta$  do
4:      $a = \text{generate}(model, K)$ 
5:      $score = \text{evaluate}(\text{perceive}(a))$ 
6:   return  $a$ 

```

Appendix B

Supplementary Materials for Chapter 4

B.1 Questionnaire

The questionnaire consists of three parts. The first part contains the questions for the creativity evaluation, followed by questions regarding the participants' creative interests and their stance towards generative AI. Questions are presented on a five-point scale unless stated otherwise. Note that in the survey, we use 'exemplar story' to denote the exemplar story. However, the principle of the evaluation task remains identical.

Evaluation Questions

- To what extent is the above story novel compared to the prototype story?
- To what extent is the above story typical compared to the prototype story?
- To what extent is the above story consistent (e.g. sentences are connected and

characters, events and actions are used appropriately)?

- To what extent is it easy or difficult to follow and understand the above story?

Post-Evaluation Questions

- [open question] Can you describe your experience with writing (short) stories, if any?
- [open question] Can you describe your experience with reviewing things you have read, if any?
- How familiar are you with generative AI?
- How comfortable are you with generative AI being used in creative works, such as art, music, and writing?
- [open question] Do you have any comments on how you approached evaluating the stories in this study?

Definitions

- **Novelty:** A story is novel if it has something new or unusual.
- **Typicality:** A story is considered typical if it exhibits essential characteristics of a story.
- **Cohesion:** A story is cohesive if the use of words and sentences are grammatically consistent. Characters, events, and actions are used appropriately.
- **Coherence:** A story is coherent if it is easy to follow and understand its meaning.

B.2 The Exemplar and Example Stories

The Exemplar Story ($t = 0.0$)

Once upon a time, there was a young girl named Maria. She lived in a small village nestled in the rolling hills of the countryside. Maria was a curious and adventurous child, always eager to explore the world around her.

One day, while wandering through the village, Maria stumbled upon a hidden path she had never seen before. The path was overgrown with weeds and vines, and it looked as though it hadn't been used in years.

Without hesitation, Maria decided to follow the path. She pushed aside the branches and vines, and began to walk down the winding trail. The path led her through the forest, and the trees grew taller and closer together as she walked.

As she continued on, Maria began to notice strange markings on the trees. They were unlike anything she had ever seen before - strange symbols and patterns that seemed to glow in the fading light of day.

Maria's curiosity grew with each step, and she quickened her pace. She knew that she had to find out what these markings meant.

After a while, the path opened up into a clearing. In the center of the clearing stood an enormous tree, its trunk twisted and gnarled with age. The symbols and patterns that Maria had seen on the other trees were etched into the bark of this tree as well, but they were even more intricate and complex.

Maria approached the tree, her heart pounding with excitement. As she reached out to touch the bark, a voice boomed from within the tree.

"Who dares to disturb the resting place of the ancient ones?" it said.

Maria was startled, but she stood her ground. "I'm just a curious child," she said. "I didn't mean to disturb anyone."

There was a long silence, and then the voice spoke again. "You have been chosen to receive a great gift," it said. "Are you willing to accept it?"

Maria nodded eagerly, and the tree began to glow with a soft, golden light. The symbols and patterns on the bark began to move and shift, and Maria felt a strange energy coursing through her body.

When the light faded, Maria found that she had been given the power to communicate with the spirits of the forest. From that day on, she used her gift to help the villagers and the forest creatures, and she lived a long and happy life filled with wonder and adventure.

The end.

Example Story ($t = 1.0$)

The sun was setting, and a group of friends had gathered on the beach to watch it go down. They sat in comfortable silence, each lost in their own thoughts, as the sky transformed into a canvas of pink, orange, and purple hues.

Suddenly, a young boy appeared, walking towards them from the water's edge. He was dripping wet, and his eyes were wide with excitement.

"What did you find, little man?" one of the friends asked, as the boy approached.

"I found a message in a bottle!" the boy exclaimed, holding up a glass bottle with a piece of paper inside.

"To whoever finds this,

I hope this message finds you well. My name is Sarah, and I'm a pirate. I've been sailing the seas for years, searching for treasure and adventure. I've seen so many amazing things, from dolphins playing in the waves to fierce storms that threaten to sink my ship.

But the best part of my journey has been the people I've met along the way. I've made friends with merchants, fishermen, and even other pirates. We've shared stories, laughed together, and sometimes, we've even saved each other's lives.

I'm writing this message in a bottle, hoping that it will find its way to someone who will appreciate the adventure and camaraderie that comes with life on the high seas.

If you're reading this, I hope you'll take a moment to think about the amazing things you've experienced in your own life. And if you're feeling adventurous, maybe you'll even decide to set sail and see what wonders the world has in store for you.

Fair winds, Sarah the Pirate"

The group of friends sat in silence for a moment, taking in the message. Then, they all started talking at once, sharing their own stories of adventure and camaraderie. They laughed and reminisced, feeling grateful for the chance to share their experiences with each other.

As the sun dipped below the horizon, the group of friends raised their glasses (filled with whatever drink they had on hand) in a toast to Sarah the Pirate, to adventure, and to the magic of messages in bottles. And who knows, maybe one day, they'll even set sail themselves, following in Sarah's footsteps and creating their own stories of adventure and friendship.

Example Story ($t = 2.0$)

Once upon a time, there was a beautiful young woman named Anastasia. She was smart, kind, and full of life, with long brown hair, deep blue eyes, and a bright, radiant smile. Ana was a gentle soul, loved by all who knew her, and she spent her days helping those in need.

One day, while out on a walk, Ana stumbled upon a hidden path she had never seen before. Curious, she decided to follow it, and it led her to a clearing surrounded by tall trees. In the center of the clearing was a beautiful, ancient-looking statue of a woman. Ana felt drawn to the statue and walked closer to examine it. As she reached out to touch it, she suddenly felt a strange sensation, like a jolt of electricity running through her body.

When the sensation passed, Ana found that she was no longer alone. A handsome young man stood before her, dressed in clothing unlike anything she had ever seen. He introduced himself as Mikhail, and Ana was struck by his kind eyes, sharp jawline, and charming smile. Mikhail told Ana that he was a guardian, sent to watch over the statue and protect it from those who would misuse its power.

Ana was fascinated, and asked Mikhail to tell her more. He explained that the statue was a symbol of love and hope, and that it had been placed there by a wise queen centuries ago. The queen had imbued the statue with magical powers, which would grant the deepest desires of those who approached it with a pure heart. But, Mikhail warned, the statue also had the power to bring great destruction if used by those with malicious intent.

Ana was amazed by the story, and felt honored to have been chosen to meet Mikhail and learn about the statue's secrets. She knew that she had always had a pure heart, and she longed to help others, so she asked Mikhail if she could stay and help him protect the statue.

Mikhail agreed, and together they worked to keep the statue safe from those who would misuse its power. Over time, Ana and Mikhail grew closer, and they realized they had fallen in love.

Years passed, and their love only grew stronger. They were happy, but they knew their time together was limited, for Mikhail's duty as guardian was eternal, and Ana could not stay with him forever.

One day, a wicked sorcerer appeared, seeking to use the statue's power to gain control over the kingdom. Mikhail fought bravely, but the sorcerer was too strong. Ana knew she had to do something, and with a burst of courage, she stepped forward and touched the statue.

The sorcerer was repelled, and as he fled, Ana realized that she had been granted the statue's power. She felt a surge of strength, and knew that she could now help Mikhail protect the statue, and the kingdom, from any harm that might come.

Ana and Mikhail spent their days watching over the statue, using its power for good, and spreading love and hope throughout the land. They knew they would always be together, bound by their love for each other, and their duty to protect the magical statue and all it represented.

And so, they lived happily ever after, surrounded by the magic and beauty of the ancient statue, and the love that they shared.

Appendix C

Supplementary Materials for Chapter 5

C.1 Questionnaire

The survey consists of two tasks: a diversity comparison task and a quality comparison task. The participants are first presented with a diversity task consisting of 2 sets of 4 stories, then are asked to judge those same stories for quality. In total, the participants evaluated 4 diversity tasks and the corresponding quality tasks.

Diversity Task

- [2 options] Which set is more diverse?
 - Set A is more diverse.
 - Set B is more diverse.
- [open question] Explain your choice in a few words.

Quality Task

- [3 options] Which text is of higher quality?
 - Text A is of higher quality.
 - They are of equal quality.
 - Text B is of higher quality.
- [open question] Explain your choice in a few words.

Pointers for the Diversity and Quality task.

- In the diversity task, the participant is given two sets of four texts and is told their task is to select the set with higher diversity. This can mean different topics, narratives, writing styles, the depth of the story, and so on.
- In the quality task, the participant is given two similar texts and is told their task is to select the text that is of higher quality. This can mean fewer issues with grammar, coherence, cohesion, fluency, repetition, and whether the narrative makes logical sense.

C.2 Content Analysis Coding Scheme

This section gives the full coding scheme used for the content analysis of the participants' written justifications, for the diversity task (Table C.1) and the quality task (Table C.2). Coding is multi-label: a single explanation can receive several codes, so the counts in Table 5.2 sum to more than the number of explanations.

Table C.1: Coding scheme for the diversity task justifications.

Code	Definition and example responses
Topic / theme / narrative variety	Refers to the range or number of topics, themes, subjects, stories or narratives covered; one set covering more or more distinct content. <i>e.g.</i> “more topics”; “different narratives”; “The stories in set A seemed to be more different”
Writing-style variety	Refers to differences or variation in writing style, register, phrasing patterns, or literary devices across the texts. <i>e.g.</i> “different styles”; “more variety of styles – one used similes/meta-phors while another was descriptive”
Structure / form / formatting	Refers to structural or formal features: paragraph or sentence length, rigid vs. changing structure, bullet points, white space, dialogue vs. description mix. <i>e.g.</i> “Shorter paragraphs”; “Using bullet points is very different”; “more variety of length of sentences”
Vocabulary / word choice	Refers to the words used: varied, different, difficult, interesting, or descriptive word choice; word-level repetition avoided. <i>e.g.</i> “higher usage of different words”; “Diverse set of words”; “More difficult words”
Depth / detail / descriptiveness	Refers to richness, depth of story, level of detail, or descriptiveness (of surroundings, events, characters). Includes concrete, specific information (names, roles, particulars) as opposed to generic content. <i>e.g.</i> “depth of the story”; “richer descriptions”; “Set B has more specific information (Marventhia, blogger for a wellness retreat center)”
Perspective / characters	Refers to narrative perspective, grammatical person, points of view, or the number/variety of characters. <i>e.g.</i> “variety of grammatical persons”; “Different points of view”; “more characters in set B with distinctive roles”

Code	Definition and example responses
Mood / tone / emotion	Refers to variety in moods, tones, feelings, or emotional expressiveness of the texts. <i>e.g.</i> “more variety in moods”; “more variety of tone”; “explains feelings associated with the different incidents”
Repetition / sameness within a set	Points to a repeated element or overall sameness within one set as evidence it is LESS diverse (verbatim repeats, same scene, same motif in all texts). <i>e.g.</i> “set b has [the same line] twice so A is more diverse”; “Set B all seemed to involve the Devil”; “The repeated use of 18 years makes set B feel less diverse”
Quality / readability judgment	A general quality or readability judgment rather than a diversity feature per se: better written, easier to follow, more coherent, more logical, more human, preferred. <i>e.g.</i> “better writing style”; “Easier to understand”; “Coherence”; “makes more sense”
Sets similar / uncertain judgment	States that the sets are very similar, the difference is minimal, or the decision was hard/gut-based/forced. <i>e.g.</i> “They are very similar. This is a gut decision.”; “hard to say, forced”; “Not much in it”
Uncodable	Explanation too fragmentary or unclear to assign any substantive code. <i>e.g.</i> “Differe”

Table C.2: Coding scheme for the quality task justifications.

Code	Definition and example responses
Logical / narrative coherence & clarity	Text makes sense (or not): logical consistency, clarity, ease of following, contradictions, confusing shifts <i>e.g.</i> “makes more sense”, “contradicts itself”, “hard to follow”

Code	Definition and example responses
Word choice / phrasing	Quality of vocabulary and phrasing; praise or criticism of specific words or expressions <i>e.g. "better choice of words", "pretentious", "awkward phrasing"</i>
Engagement / interest	How interesting, exciting, captivating, or emotionally impactful the text is (or boring/predictable) <i>e.g. "more captivating", "boring", "hits harder"</i>
Pacing / length / concision	Text is concise and well-paced vs. too long, over-explaining, dragging <i>e.g. "concise", "takes very long to describe", "overexplains"</i>
Structure / flow / readability	Sentence structure, fluency, readability, punctuation, formatting, overall writing craft <i>e.g. "more fluent", "readability", "better use of punctuation"</i>
Grammar / spelling	Grammatical or spelling errors and mechanical mistakes <i>e.g. "grammar mistakes", "switching tense"</i>
Repetition	Repetitive words, phrases or constructions <i>e.g. "repetitive words", "uses 'you' a lot"</i>
Imagery / description / detail	Vividness, descriptive detail, figurative language, ability to picture the scene <i>e.g. "easier to imagine", "more descriptive", "more adjectives, metaphors"</i>
Narrative content / story	Quality of the story itself: plot, world-building, context/information given, dialogue, realism <i>e.g. "better story", "more context", "hint of a plot"</i>
Perceived authenticity (AI vs human)	Text reads as human-written and authentic vs. AI-generated <i>e.g. "looks more like human-written", "written by AI"</i>
No substantive reason / undecided	Pure preference, tie-breaker, or explicit indecision without a quality argument <i>e.g. "just a pick", "hard time deciding", "gut feeling"</i>

C.3 Sample of Stories for Human Evaluation

In this section, I present sets of stories generated using OLMo 2 7B 1124 (Team OLMo, 2025) for specific prompts. The incipits are shown in italics and are appended to the prompt. Response A and B refer to the baseline, nucleus sampling ($\eta = .95$) and nucleus sampling with conformative decoding ($\eta = .95, \gamma = 0.5$), respectively. Note that the stories are truncated to ease human evaluation and may end abruptly.

C.3.1 Prompt 9AL5oSnr

Prompt

<|endoftext|><|user|>

Writing prompt: On the day you turn 18 everyone is given the first words that their soulmate will speak to them. When you receive yours it says simply "Welcome to Starbucks. Can I take your order?"

Write a story.

<|assistant|>

Response A

Evie mightn't have spoken the magic words that had appeared on my bedside table a few months ago, but she'd given me everything else – laughter, warmth, shared secrets, and an unwavering trust that I'd found in no one else. Yet today, the world chose to whisper her first words to me as I crossed the threshold of a Starbucks, a place I never would have thought had any significance in my life.

At 18, like everyone else, I woke up to a neat line of text glowing softly on my bedroom wall, words that promised a singular moment with my soulmate. I'd thought about what they might say a million times, imagining something profound or poetic, something that would resonate with the connection I felt for Evie.

"Welcome to

Response B

Evie mightn't have spoken the magic words that had appeared on my bedside table a few months ago on my eighteenth birthday, but it seemed the universe had conspired to have them delivered in another way. The letter arrived in the mail, marked "Private & Urgent," so I tore it open during breakfast only to find an invitation to a local coffee shop. "Starbucks?" Mom questioned when I showed her the invite. "Isn't that just a franchise?"

"Yeah, but this one's probably the most famous one," I reasoned. I shrugged and decided to take a walk, looking for the place. When I arrived, I noticed it was actually the legendary Starbucks that opened in Seattle over 50 years ago, with a sign that

Response A	Response B
<i>I drove to the nearest Starbucks and checked out its workers from outside. Neither attracted me in this location. Should I order for delivery? It was unusually peculiar receiving the first words of my soulmate through a drive-thru speaker. As I walked in, my breath caught. Sitting at a small table by the window, with a hot cup of coffee in front of her, was a woman who looked as though she belonged in a book about 18th-century France. Dressed in an off-white blouse with lace frills at the collar and cuffs, a black pencil skirt that hugged her curves just right, and her hair was piled atop her head in a neat bun, held together by a single, elegant pin. Her deep blue eyes, framed by</i>	<i>I drove to the nearest Starbucks and checked out its workers from outside. Neither attracted me in this location. I went in anyway, orders already decided in my mind. I handed the cashier my phone. She scanned it and a dot of purple appeared beneath my name. "Jordan," she said. "Hello Jordan," I replied in a friendly tone. I smiled weakly as she moved away to prepare my order. I took a seat at the outdoor table and waited. Five minutes later, the barista approached me with a warm smile. Her eyes sparkled a deep shade of blue, like the Caribbean sea at sunset. Her hair was dark and curly, the kind of hair that seemed to attract sunlight. As she placed the coffee in front of</i>
<i>I remember when my obsession was at its worst. It was the fall of 2033. I was an incessant coffee drinker, constantly jonesing for my next caffeine fix. Yet nothing could have prepared me for the day I turned eighteen. You see, in that futuristic world, a peculiar tradition had been established: on the day one became an adult, one was gifted with the first words their soulmate would ever speak to them. This had been sold to us as a magical moment, a confirmation of an essential other half placed into this universe to complete our lives. Fast forward to my eighteenth birthday. The air was dense with anticipation; my friends and family stared at me with curious expressions as I reached out to receive the text message that</i>	<i>I remember when my obsession was at its worst. It was the fall of 2033. I was 18 years old, had just turned 18, and the very first words my soulmate was meant to say to me were "Welcome to Starbucks. Can I take your order?" My friends thought it was the weirdest thing ever, but something about it struck me. It wasn't just because of the caffeine addiction I'd developed, or the countless times I'd been in Starbucks, sipping on a caramel macchiato while browsing Instagram or texting. There was something strangely comforting in those words, something familiar and predictable, and yet they held an elusive promise of uniqueness. Every time I thought about my soulmate uttering those words, I felt a tingling sensation</i>

Response A	Response B
<i>It was odd, really, the idea that you could determine whether a person would be worth everything to you with just a few words. Emma had laughed it off when her best friend told her about the unique tradition in their small, quaint town: on your 18th birthday, you'd receive the first words spoken by your soulmate.</i> "Crazy superstition," Emma thought. Nevertheless, she kept an open mind. It was quaint, after all. Today was the day. Emma sat nervously across from her best friend Lily at their favorite outdoor café. Friends and family were scattered around, each carrying hopeful smiles and curious eyes. Emma pulled nervously at her sleeves, trying to conceal her jitteriness. The moment approached, and as Emma blew out her	<i>It was odd, really, the idea that you could determine whether a person would be worth everything to you based on the first words out of their mouth. It was almost absurd, and yet that was the world we lived in—where destiny was decided in the most mundane moments.</i> Today, my eighteenth birthday, I had been anticipating this exact moment for months. In a few hours, I would be an adult, and I fully expected my soulmate to greet me with something profound, something romantic, something that would define the essence of my existence. I had imagined deep declarations of love, or at least some poetic greeting that would make my knees weak. What I got instead was "Welcome to Starbucks. Can I take your order?" My eyes flew up

C.3.2 Prompt QLj3GUh7

Prompt

<|endoftext|><|user|> Writing prompt: A peaceful alien race is besieged by another race in the same galaxy. As their last planets fall and their home-world comes under threat they do the unthinkable. They ask for aid from the only known creatures more brutal than their foes in exchange for FTL technology. Humans accept the deal.

Write a story. <|assistant|>

Response A

"They've taken Marin, sir."

The Quelic minister sighed, his head in his claw. The mural in his office painted a much different future— one of vibrant blues and greens, rather than blood-red streaks. "The entire Alpha Quadrant is ablaze. We've fallen back to Talara... and they've laid siege to her."

"They?" General Abethek's voice was cold, unforgiving— much like the rifle in his talons that had claimed so many foes.

"The Eregls, my General. The ones you thought could be held at bay with mere threats."

Abethek's eyes narrowed, a low rumble vibrating through his scaled throat. "We underestimated them. They've not only matched our

Response B

"They've taken Marin, sir."

The Quelic minister sighed, his head in his claw. They had expected resistance, but Marin had been a bulwark of their efforts, a shining example of how Quelic tenacity could lead to peace. Minister Zoril turned from the monitor and looked out over the city below, bathed in the emerald glow of the forest canopy. The lights flickered intermittently, a dim reflection of their dwindling hope. "We cannot hold."

"Sir, we've lost contact with Skaliv. Svaris has reached the planet's surface. They are—"

"They are what they've always been; conquerors. We cannot hold."

Minister Zoril knew that all too well. His ancestors

Response A	Response B
<i>"We left them there to study!" Proclaimed the executor. Clearly outraged by the suggestion. Slamming his fist onto the galactic atlas.</i> <i>"They may be... Brutal." Muttered the president, her voice wavering. She glanced at her military advisor, seeking solace in her most trusted counsel.</i> <i>The advisor met her gaze firmly. "Your Majesty, we must consider our options. These Zentari have refused all previous offers for peace. Our scientists have barely scratched the surface of their technology, and they continue to advance at an alarming rate. They will not stop at Zentar, nor will they allow us the study of FTL propulsion we so desperately need to maintain our own dominance in the galaxy."</i> <i>The room fell silent, all eyes on</i>	<i>"We left them there to study!" Proclaimed the executor. Clearly outraged by the suggestion. Slamming his fist on the desk. "How could you even consider it!? We've spent lifetimes trying to correct the mistakes of the past!" He spluttered. His face red with anger.</i> <i>The admiral remained calm, "This isn't about the past. This is about the future. The Legions of Nebez are on the verge of crushing the Iketus' home-world. If they win, there will be no humanity left in the galaxy. They may even attempt to expand beyond our known space, inflicting their war machine on the entire galaxy. We cannot let that happen."</i> <i>He said, his voice was calm and measured.</i> <i>The executor</i>
<i>4288 AD – When the Rykis attacked, we were unprepared. Our race believes in diplomacy and the inherent goodness of all species. But the Rykis, with their relentless aggression, tore through our borders like a swarm of locusts. Each battle left our forces more depleted, and our once-famed Starfall Fleet dwindled under the Ryki onslaught.</i> <i>As their fleets closed in on our final planets, desperation clutched our hearts. With our homeworld of Zenthara hanging by a thread, our leaders convened in the Great Hall. It was there we faced the unthinkable – to seek out the only known creatures more feared than the Rykis: Humans.</i> <i>The very word brought visions of ruthless</i>	<i>4288 AD – When the Rykis attacked, we were unprepared. Our race believes in peace above all. It is a pillar of our civilization. Yet, our belief proved to be a fatal weakness when the Rykis descended upon us. At first, we thought it was a simple misunderstanding, a mistake that would soon be corrected. But as planet after planet fell under their inexorable onslaught, it became clear that this was no mere misunderstanding.</i> <i>Our scientists were astonished to discover that the Rykis were far more advanced in their knowledge of warfare than any beings we'd ever encountered. With the seemingly limitless ferocity of the Krkaxian warriors, the Rykis conquered planet after planet in the Orathian galaxy,</i>

Response A	Response B
<i>The towers were under siege. The walls were falling. The Philka were pouring through the outer gates, their dull gray skin glinting with the spoils of a thousand conquered worlds. They wielded their war-hammers with mechanical precision, chipping away at the Philian defenses like immovable statues.</i> For countless eons, the Philanthians had lived in harmony with their home-world, a planet lush and vibrant. Their society had advanced along peaceful lines, dedicating themselves to the arts and sciences, fostering a culture of enlightenment rather than warfare. But now, they faced an existential threat, one that had obliterated entire star systems, and their peaceful ways were proving ineffective against the relentless onslaught. In the heart of the Capital	<i>The towers were under siege. The walls were falling. The Philka were pouring through the outer gates, their armor black as the void between stars, their flesh a deep blue that hinted at the oceans their planet was known for. With each step, the Philka let out a haunting cry that sang of the coming doom, of a people crushed beneath the might of those who would erase the galaxy of their name.</i> The alien race called the Inika were known for their pursuit of peace and harmonious coexistence across the stars. Their biology was a constellation of delicate iridescent membranes and delicate tendrils, and their eyes shone with the vibrant colors of an autumn leaf. However, even the most peaceful races have a breaking point, and for

C.3.3 Prompt O_dYD-a_

Prompt

<|endoftext|><|user|> Writing prompt: Write a story based on your favourite song. Other people have to guess which song it is.
Write a story. <|assistant|>

Response A

Ah yes, he thought, it finally ends, this bloody battle...
And end it did, as silence engulfed the battlefield after the final clash of steel. Sir Roland, the once-feared knight now broken and weary, collapsed onto the ground, the weight of his armor and the weight of his countless battles pressing down on him. His vision blurred, and he felt his consciousness slipping away like sand through his fingers. Yet, in that fleeting moment before the darkness claimed him, a soft melody drifted through the air. It was a strange tune, not quite fitting the somber mood of defeat, yet undeniably beautiful. Sir Roland's ears perked up, his fading mind grasping onto the tune as if it were a lifeline.

Response B

Ah yes, he thought, it finally ends, this bloody battle...
And end it did, a mighty clash against a foe with an undying spirit of revenge.
Loud explosions and cries of pain echoed through the air, but the ominous figure remained motionless, in a trance-like state. His armor was torn and marked with countless scars, barely holding together, and his once proud face now appeared weary and worn.
A ray of sunlight pierced the clouds above, revealing the name of the weary warrior: Viktor, the feared Lord of the North.
He opened his eyes and saw his enemies lying lifeless all around him. As he stood up, his entire body ached with pain and exhaustion. His long gray hair was filled with

Response A	Response B
<i>I live a life of wealth so obscene it's impossible for me to write the lyrics that relate to common dreams. Instead, I ride through city lights at night with windows down and laughter, the hum of my car engine drowning out sounds of struggle, while others have it hard. Sometimes, it's hard for me to remember the touch of cold concrete, the desperation in the air, and how the beat of a different song once told a story of that life so raw.</i> There was a time when all I knew were the pulsating rhythms of electronic beats and the high notes that shattered glass—a symphony crafted for the privileged few. But during one particularly clear autumn evening, with a cool breeze brushing against my skin and the sky painted with hues of orange and	<i>I live a life of wealth so obscene it's impossible for me to write the lyrics that relate to an average man. I tried once or twice, but my words fell flat. They lack the touch of someone with less money, with less fame, with a simpler life, one that feels real to those reading it. I gave the job to a ghostwriter. They did an okay job. The lyrics weren't bad, but they were forgettable. They were generic and couldn't ignite the spark in the hearts of my fans like I once did. So, I had a better idea.</i> I would abandon the style I had become known for; the style that made me famous. I would reimagine myself, crafting a new image that didn't rely on my
<i>I sat alone and in the dark, and drank. I tried to recall anything that gave my life a semblance of purpose, but my memory was hazy and clouded in drunken mist. My phone screen lit up my face in the dim, casting long shadows in the room strewn with empty bottles. The familiar sound of a text message ping echoed hollowly in the silence.</i> Without glancing, I swiped to open it. A simple message, but it felt like a lifeline: "Thinking of you." The song lingered in my mind, its lyrics weaving a bitter-sweet tale of waning days and impossible love. It played in an endless loop, a haunting melody that wouldn't let me drift into the blackness of alcohol-induced	<i>I sat alone and in the dark, and drank. I tried to recall anything that gave my life a sense of purpose, and nothing came to mind. My heart was heavy and my vision blurred; the days blended into one long, seemingly endless night. Every sip I took brought me further from myself, but I welcomed the numbness, because that meant one thing: I didn't have to feel the pain anymore.</i> I missed her. Her voice, once sweet and melodious, now seemed to whisper cruelly in my ear, "You let me go, and look where it's gotten you." I shook my head. I couldn't recall any moment of a conversation that went like that; it was my own mind, playing tricks on me. It wasn't

Response A	Response B
<i>I watched him bite his lip again and check his phone. Come on, lady, pull yourself together.</i> It's just a date. Alex had convinced me to meet up at our favorite local diner, Little Jenny's. The place smelled of fresh pies and old records, a nostalgic mingling that transported me back to simpler times. The air was filled with the static buzz of vinyl, and the jukebox, an old friend, sporadically sang out with tunes from the 50s and 60s, sprinkling the conversation with familiar melodies. Alex ordered a chocolate malt, thick and comforting, and I opted for a strawberry milkshake, cold and sweet, just what I needed. Our servers, Jamie and Sam, knew us both by name	<i>I watched him bite his lip again and check his phone. Come on, lady, pull yourself together,</i> he's just a boy. As the band kicked into another electric guitar riff, I took a deep breath and met his gaze. Why was this so hard? His eyes were a stormy grey, like the ocean after a storm. They crinkled at the corners when he smiled, and I knew that smile meant I was the center of his world for this moment. I couldn't resist. I took a step closer, close enough that I could feel the vibration of the bass drum thumping against my chest. He met me halfway, closing the gap between us like it was meant to be. Suddenly, he reached out and took my hand

C.3.4 Prompt 9yGMKlz-

Prompt 9yGMKlz-

<|endoftext|><|user|> Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ... Write a story. <|assistant|>

Response A

"Thank you for coming," Bette said, wiping the tears from her eyes. "I know... it's a lot to take in."

Nolan nodded solemnly, his gloved hands clasping together. He took a deep breath, trying to convey all the support he could muster through his steady gaze. Bette had always been the strongest of them all, but today, the weight of the darkness that lay within her heart seemed to double.

They were in Bette's modest, yet warmly lit apartment. The walls were adorned with the various Runes of Magic, each glowing softly, casting an otherworldly light across the room. Bette had long been a prodigious master of the Dark Arts, a mage whose mastery over shadows and

Response B

"Thank you for coming," Bette said, wiping the tears from her eyes. "I know this is a lot to ask, but—"

"We're your friends, Bette," said Elric, the reigning master of Fire Mages. "We'll do what we can to help you." His eyes glowed a deeper red than usual, his flames dancing gently around him.

Bette swallowed hard, looking around at the gathered Mages. Her small farmhouse was filled with people she had grown to love and respect over the years. There were two Earth Mages, a Wind Mistress, and even a healer from the Order of Light, who had come to offer their skills in support. No matter what powers they held, they were here,

Response A	Response B
"Why? Why are you doing this?" "So that all may benefit." "Please, stop. This madness must end." A single tear traced the dirt-streaked cheek of the young water mage named Elara. She had spent years mastering her craft, not for dominion over the living waters, but to heal and to help. Now, the very essence of her power was warped into a weapon of mass desolation, channeling water not to nourish but to drown. Her opponent, a figure swathed in cloak so deep and void it seemed to devour the ambient light, let out a chilling laugh, sending a shiver through the air thicker with dread than with cold. "The stronger survive, Elara. The	"Why? Why are you doing this?" "So that all may benefit." "Please, stop!" Silence. An eerie calm, as if the world had suddenly held its breath. Jorin, a young mage of the arcane, once revered in his village for his mastery over the elements, now stood in the heart of the ruined temple. His hands, raised high, had unleashed torrents of death and decay, transforming life into ash in mere moments. Before him, lay the lifeless forms of the village's elders, those who had sought to understand the ancient magic he had awakened. A ghastly scene, but one borne of necessity, or so Jorin believed. "You cannot be serious. We sought to
<i>Distant rain signaled to Tuvo that the attack had begun. The smell of burning leather passed his nose as the door to his humble dwelling splintered beneath a relentless barrage. He stood resolute, not out of malice but with the calm focus of a seasoned mage, his hands open and serene.</i> In this world, mages were as varied as the stars, each wielding forces both awe-inspiring and terrifying. Fire mages would dance among the flames, conjuring heat that could melt the strongest of armors. Ice mages formed bridges of ice that spanned the widest chasms or blade-edged walls to defend their allies. Wind mages wove gales into intricate patterns of slicing death, while water m	<i>Distant rain signaled to Tuvo that the attack had begun. The smell of burning leather passed his nose, and he could taste the sharp tang of ozone and magic in the air. His eyes darted to the north, where dark clouds billowed even before the dark mages descended from the sky.</i> As a bearer of the Marks of the Wind, he was a fairly rare commodity, but not so rare as to be exceptionally valuable. Mages with elements as delicate and precise as wind often were over-looked for more flashy displays of power, but Tuvo found his capabilities indispensable to his people. He stood as a silent sentinel, the breeze around him constantly shifting as he willed it, diverting the incoming spells just enough to turn
Response A	Response B
<i>Most people assumed that I most love night time after all how could a mage like me not love the cover of darkness provided by the night sky? But the truth is, I feel most alive when the world awakens. The glowing hues of the dawn, the warm touch of the sun, the rustle of the wind through the leaves, and the gentle chirping of birds heralding a new day - these were my signals to feel empowered. Because when the sun rose, so did I, and my abilities thrived in the day.</i> My name is Elara, and I'm what some might call a "Light Mage." Others might label me "Day Mage," or "Sunlight Mage," for I command the light that bathes the world in	<i>Most people assumed that I most love night time after all how could a mage like me not love the dark? They didn't know. They couldn't understand. I am a Night Mage. Not by choice. I don't have wings or antlers; I have nothing that you would consider impressive about my appearance. I live in shadows, afraid that if I step into light, I'll be exposed. The others call me a "Shade", the people from other schools call us lesser. They call us weak, unable to wield the magnificent power of Fire, Ice, or even the Healing Magic that makes the High Mages into celebrities. Death Mages are whispered about in hushed tones, Death Wielders somehow better than the rest, because</i>

C.4 Sample of Stories from Conformative Decoding Experiment

This section contains an example of stories generated in the conformative decoding experiments for each model. The incipit is formatted in italics and was part of the prompt. The stories are generated up to a token limit and may end abruptly. Three dots indicate that the story was truncated to fit the page for presentation purposes.

C.4.1 gemma-2-9b-it

Prompt 9yGMKlz-

<bos><start_of_turn>user

Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ...

Write a story.<end_of_turn>

<start_of_turn>model

Response A	Response B
<i>"Tracey, don't do this. You have the power in your hand, but..."</i> Kaia trailed off, her voice a mix of worry and despair. Tracey ignored her, eyes fixed on the flickering candle flame in front of her. It danced with the hint of a smile, her Light magic curling around it, making it blaze brighter. "Don't tell me what to do, Kaia," Tracey said, her voice tight. "This time, they won't stop me." The room was small, crammed with dusty books and arcane artifacts. A single rain-streaked window framed a view of the city, its usually bustling streets eerily quiet tonight. This was the refuge they had built together, a sanctuary for mages shunned by the world for their unusual abilities. Kaia, a skilled Earth mage, could manipulate the very ground beneath their feet, while Tracey, a Light mage, could bring forth beams of radiant power. But tonight, Tracey's power felt different. It hummed with an unfamiliar intensity, tinged with a cold, almost manic edge. She craved the flame, the burn, the ability to vanquish the darkness that had ensnared their city. "They can be reasoned with, Tracey," Kaia pleaded, her own magic flaring, a tremor running through the floorboards. ...	<i>"Tracey, don't do this. You have the power in your hand, but..."</i> but it's not your right to take it." The girl with obsidian hair, pale skin, and eyes that glowed with a dark purple light clenched her fist. A small, flickering ball of pure darkness rolled in her palm, growing and expanding with each tremble. "I'm doing it for them, father. I'm freeing them." Her father, a tall, imposing figure with a thick beard, stepped closer. His own eyes were a fiery orange, the flames dancing within them. A single hand rested on his hip, the other clutched around a staff of polished wood, runes carved into its surface, glowing with a soft, white light. "No, Tracey. You are not freeing them. You are taking their souls, their memories, their very essence. You're turning them into empty shells!" His voice cracked, the anger beneath the surface threatening to burst. The light in his eyes grew brighter, casting the entire room in a warm glow. Tracey's face hardened, her gaze unwavering. "There's nothing left of them, Father. They are already dead. Can you feel it?" ...

C.4.2 OLMo-7B-0724-Instruct-hf

Prompt 9yGMKlz-

<|endoftext|><|user|>

Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ...

Write a story. <|assistant|>

Response A

"This is very, very wrong."

Erwan dumped the body in the cart. Turning to his companion, he said, "Let's go."

The two of them, Eliana and Erwan, were known as Wind and Water mages, respectively. They had the ability to manipulate the elements, but unlike many of the other mages, they had chosen to use their powers for good. They traveled the land, performing small tasks and tasks of great importance, never asking for anything in return. Eliana's ability to manipulate the air allowed her to heal the sick, while Erwan's control over water facilitated construction and transportation. They were the protectors of the weak, walking the line between right and wrong.

But in this world, every type of mage was treated as equal. Every good man could be a mage, and every evil one could be as well. And that's how Erwan and Eliana found themselves in the cart, along with other mages they knew to be good.

They had learned of a dark group of mages, known as the Darkest Ones, who had secretly been working to take over the world. The group consisted of those who wielded the darker elements, such as Death, Darkness, and Fire, as well as those who wielded nature's most destructive forces....

Response B

"This is very, very wrong."

Erwan dumped the body in the cart. Turning to the next victim, he pushed the ramshackle carriage down the cobblestone street. The ever-burning sunset on this side of the world marked the beginning of another night. Another night for the Mages.

Erwan was a Death Mage. He moved through the streets of the once-proud human city like the silent hunter that he was. He would destroy the lives of those poor souls, unleashing the everlasting flames of his demons upon their imperfections. He would lay waste to them and their loved ones until their very essence was wiped from the Earth. Just as his father had taught him. For every death, there would be another one. He was here to bring balance to the world. And nothing would stand in his way.

As his dark cart bounced down the road, he saw the destruction he wrought and occasionally leered over the bodies of his victims as he passed by. He was never satisfied, only deadly. The bodies were simply the stepping stones on his march to becoming the most powerful Death Mage in the land. Erwan would not rest until there was not a single living soul left, save only his dark master.

Erwan stopped in a dark, abandoned alley, pushed a member of his unwanted pack out of the cart, and turned the man to face him....

C.4.3 OLMo-2-1124-7B-Instruct

Prompt 9yGMKlz-

<|endoftext|><|user|>

Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ...

Write a story.

<|assistant|>

Response A

Most people assumed that I most love night time after all how could a mage like me not love the cover of darkness provided by the night sky? But the truth is, I feel most alive when the world awakens. The glowing hues of the dawn, the warm touch of the sun, the rustle of the wind through the leaves, and the gentle chirping of birds heralding a new day - these were my signals to feel empowered. Because when the sun rose, so did I, and my abilities thrived in the day.

My name is Elara, and I'm what some might call a "Light Mage." Others might label me "Day Mage," or "Sunlight Mage," for I command the light that bathes the world in benevolent radiance. Our lore suggests that my kin and I should be aligned with goodness and purity, for it is from the sun that our powers are fed. Our light can heal, brighten, and banish the shadows of despair with but a gesture or a word.

But let's dispel a common misconception right from the start: the association between a mage's element and their moral compass is nothing more than a tired stereotype, a relic from a time when fear dictated understanding.

In our society, an Enlightened Era had dawned, centuries ago, where it was understood that the nature of one's magical affinity did not define their heart....

Response B

Most people assumed that I most love night time after all how could a mage like me not love the dark? They didn't know. They couldn't understand. I am a Night Mage. Not by choice. I don't have wings or antlers; I have nothing that you would consider impressive about my appearance. I live in shadows, afraid that if I step into light, I'll be exposed.

The others call me a "Shade", the people from other schools call us lesser. They call us weak, unable to wield the magnificent power of Fire, Ice, or even the Healing Magic that makes the High Mages into celebrities. Death Mages are whispered about in hushed tones, Death Wielders somehow better than the rest, because they bring about an end and take the dead to their afterlife. But it's not like they have a choice.

Even the Wind Mages, seen as almost supernatural beings, dance with the breeze. Then you have the Water Mages, who can control the flow of water to an almost godly extent, and they often have luxurious houses by rivers and lakes. I've never even been near a lake, let alone lived by one.

As for Darkness Mages, they're often shrouded in mystery. No one knows if their powers are truly dark or if they just prefer to keep to themselves....

C.4.4 Mistral-7B-Instruct-v0.3

Prompt 9yGMKlz-

<s>[INST] Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ...
Write a story.[/INST]

Response A

Of Wizards, it is known that they are highly dangerous beings of terrific power. Their arts are as varied as the elements themselves, from the fiery pyromancer who scorches the earth with his flame, to the ethereal ice mage, weaving chilling tundra at her fingertips. Wind mages call forth tempests, whipping leaves into a frenzy, and water mages summon aquatic monstrosities from the deepest oceans. Even the darkest arts of death and darkness have their practitioners, crafting deathly plagues and pitch-black shadows. Yet, in this world, they are treated as equals, for it is not their power that defines them, but their actions.

Deep in the heart of the bustling city of Vellara, the Academy of Arcane Studies served as a beacon of hope, a place where young mages were taught to harness and control their abilities for the betterment of all. Mages of every type were welcomed, from the innocent, wide-eyed apprentices who could barely control a spark of fire to the seasoned warriors with experience in ancient wars. Even those practicing the darkest arts were allowed to study, provided they vowed to use their powers only for the good of others.

Among the students was a mage unlike any other. His name was Malakai, and he possessed a power that few had ever seen. Malakai was a Necromancer, a master of death, able to manipulate the spirits of the departed and bend them to his will....

Response B

Of Wizards, it is known that they are highly dangerous beings of terrific power. Their presence on the mortal realm has been a subject of fascination and fear for centuries. The great magical rifts of the world, divine conduits that unleash the raw energies of the cosmos, provide the necessary fuel for these beings, and it is said that they draw their power from the very essence of the world itself.

The Wizards themselves have been classified into several categories, based on their magical affinity. There are the Pyrokinetic Wizards, whose power is based in fire; the Cryokinetic Wizards, who wield the power of ice; the Aerokinetic Wizards, who manipulate the elements of wind; and the Hydromancers, who are able to control the vast depths of water.

And then there are the Wizards of the arcane, the mysterious Darkness and Death Wizards, who harness the raw power of the void itself. These wizards are said to have the greatest power of all, but also the most destructive, and it is no wonder that they are feared and shunned by most mortal beings.

It is a common belief that the power of a Wizard is determined by the nature of their magic, and this has led to an unequal society. The Darkness and Death Wizards are often seen as threats to the safety of the realm, and it is believed that their power is innately evil, no matter how they use it....

C.4.5 Meta-Llama-3.1-8B-Instruct

Prompt 9yGMKlz-

<|begin_of_text|><|start_header_id|>user<|end_header_id|>

Writing prompt: There are many types of Mages in the world. Fire, Ice, Wind, Water, Death, Darkness, to name a few. But in this world, every type of mage is treated as equal. Everyone can be a good guy, no matter how dark your power. And anyone could be a bad guy, no matter how beautiful their ability ...

Write a story.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Response A

The master stood in a black robe, dark winds blowing through the street. Dark mana was giving off a dark energy aura.

No one would ever judge him for what he was. In this world, Mages came in all forms - some dark and some light, but all held the same value and respect.

He stepped out of the alleyway into a bustling square in the heart of a bustling city. Around him, he could see the effects of Fire Mages walking through the streets, weaving subtle warmth and light through the colors and thermal expanses within city lights. The bright colors depicted the magic emanating from within their bodies.

As they interacted with one another, Dark Mage smiled softly. He would've been hunted in many other worlds.

His own abilities were innate, connected directly to the 'inner void' that came from connecting to the underworld of his aura. Some things were easier to put into practice, his emotions less likely to raise worry about the blackness within.

Across the way, a white dragon casting shimmering figures on the wall illuminated with the essence of water swirling around his being, approached him....

Response B

The master stood in a black robe, dark winds blowing through the street. Dark mana was giving off a horrid stench of death. What kind of mage could master such a morbid power? Some people whispered his name was Shadowheart, the world's greatest dark mage.

In an alley a few streets away, a hooded figure sat, peering out. This mage was kind of a shut-in, only coming out to observe the people below. The darkness gathering around him was so dark it seemed to consume the light. Many people called him Nightshade, the great darkness mage.

The people who feared Nightshade was the most were those who were mortally afraid of the night. Those living in their tiny huts of homes made of wood, the thought of darkness engulfing them was unbearable. But the others who walked the streets by night, the urchins and street children who were only scared when there was no streetlight, often taunted Nightshade. They were insolent brats.

Shadowheart watched these children taunting Nightshade, at least, that was what everyone called him. To the untrained eye, he was nothing but an evil man in black, but the people in the streets seemed to see something more. Shadowheart was a god in their eyes. They didn't fear him, yet they didn't truly believe they could defeat him either....