

STAND-UP AND DELIVER:
MODELLING STAND-UP COMEDY TO ASSESS THE
SELF-IMPROVEMENT ABILITIES OF LARGE
LANGUAGE MODELS

A THESIS SUBMITTED TO
THE UNIVERSITY OF KENT
IN THE SUBJECT OF COMPUTER SCIENCE
FOR THE DEGREE
OF MSC-R.

By
Thomas Hills
June 2026

Abstract

This thesis investigates the self-improvement capabilities of Large Language Models (LLMs) through the domain of stand-up comedy. With its fundamental reliance on continuous feedback and iterative refinement, stand-up presents an ideal medium for exploring whether computational systems can demonstrate genuine creative evolution. Specifically, this research focuses on the generation of long-form anecdotes as part of a cohesive narrative routine, moving beyond the traditional computational focus on short-form wordplay.

To achieve this, the study introduces a novel prompt-engineering methodology designed to guide several commercial LLMs through a simulated "Engagement-Reflection" cognitive model. This pipeline forces the models to sequentially draft, self-critique, and iteratively improve comedy material. Throughout this process, the models demonstrate the capacity to identify structural flaws and independently apply revisions that align with established comedic mechanics, such as comedic timing, incongruity, and known techniques of the medium.

An analysis of the results reveals that clear, actionable improvements are both identified and executed by the models. By employing a triangulation evaluation strategy, the research highlights substantial quantitative and qualitative differences between the iterative drafts, showcasing measurable changes in pacing, semantic cohesion, and the underlying structure of the humour.

This thesis presents a promising framework for interacting with LLMs to reach an optimal creative result. The generated material exhibits clear "Progression

and Development”, fulfilling a key component of computational creativity, and suggests significant applications for these models within the broader creative writing industry. However, the ultimate success of the generated routines as viable stand-up material remains unclear. Determining true performative validity requires further evaluation by domain experts or through live stage performance.

Acknowledgements

I would like to thank my wonderful supervisor Anna Jordanous for her constant support and advice.

I would like to thank my supervisory chair, David Barnes, and the head of PGR studies Rogerio De Lemos, for their advice and support making it through the required checkpoints of my MSc-R.

I would also like to thank Shivali Hills, Max Peeperkorn and Ani Shaw for their advice and reading of my research at various points throughout the course.

Contents

Abstract	ii
Acknowledgements	iv
Contents	v
List of Tables	viii
List of Figures	x
List of Prompts	2
List of LLM Outputs	3
1 Introduction	6
1.1 Bringing Computer Science and Stand-Up Comedy together . . .	7
1.2 Research Questions	8
1.3 Thesis Structure	9
2 Background	10
2.1 The Art and History of Stand-Up	10
2.2 Foundational Humour Theories	15
2.3 The Mechanics of the Joke	21
2.4 The Performance Layer	27

2.5	The Process of Refinement	31
2.6	Chapter Summary	33
3	Literature Review	35
3.1	Computational Creativity Frameworks	35
3.2	LLMs and Prompting	42
3.3	AI in the domain of Humour	54
3.4	Evaluating Humour	61
3.5	Ethics in AI Comedy	69
3.6	Chapter Summary	74
4	Methodology	75
4.1	The Purpose of the Research	75
4.2	Prompt Engineering	79
4.3	Phase 1: Topic Choice (Engagement)	90
4.4	Phase 2: Generating Material (Engagement)	94
4.5	Phase 3: The Simulated Feedback Loop (Engagement) / (Reflection)	103
4.6	Phase 4: Performance Annotations (Engagement)	112
4.7	Chapter Summary	116
5	Evaluation	118
5.1	Results	118
5.2	Phase 1: Topic Choice	122
5.3	Phase 2: Material Generation	130
5.4	Phase 3: The Simulated Feedback Loop	139
5.5	Phase 4: Performance Annotations	169
5.6	Survey and External Review	177
5.7	Chapter Summary	204

6	Discussion	208
6.1	Successes of the Research	208
6.2	Limitations of the Research	213
6.3	Ethical Implications of AI-generated Comedy	222
6.4	Future Work	227
6.5	Chapter Summary	234
7	Conclusion	236
	Bibliography	239

List of Tables

1	A comparison of SMART and CLEAR prompt frameworks	52
2	Performance and Delivery Annotations	113
3	The average percentage change of each routine at each iteration of refinement.	140
4	The Normalised Levenshtein edit rates across iterations.	142
5	The average Semantic Distance by each LLM model across each iterations.	158
6	The average semantic distance in Iteration 3 and categories by LLM model.	159
7	Comparison of average Semantic Distance and categories for each experiment.	160
8	The raw number of annotations added to each generation.	170
9	The normalised score based on annotations relative to word count.	171
10	Comparative analysis of LLM model’s annotation styles and structural consistency.	173
11	The average annotation detail score for each model and the count of max scores.	174
12	Categories based on each model’s average detail score.	174
13	The percentage of ‘inspired’ or ‘developed’ unique annotations in each generation.	176

14	Categorisation and examples of the generated material across iterative stages.	178
15	Audience preference voting percentages across each iteration stage.	179
16	Wave 2: Audience preference voting percentages across each iteration stage.	181
17	Wave 1: Audience preference voting percentages broken down by individual model across each iteration stage	183
18	Wave 2: Audience preference voting percentages broken down by individual model across each iteration stage	183
19	Distribution of overall sentiment across the qualitative feedback comments provided by survey participants.	188
20	The distribution of audience-preferred jokes mapped against their Semantic Distance zones.	191
21	A cross-referenced breakdown of winning audience votes, categorized by the joke's underlying Semantic Zone and its corresponding iterative stage.	192
22	The top-performing individual jokes across the generated trios in the external survey	195
23	The winning iterations across all ten generated trios in the external survey, ranked by audience preference.	196
24	Chi-Square Goodness of Fit test results comparing observed audience preferences against an expected uniform distribution ($N = 260$). .	202

List of Figures

1	Diagram of the Benign-Violation Theory, from (Hills 2026)	19
2	Example prompt, applying concepts of ‘Prompt Chaining’, specifically ‘Generator’, ‘Critic’ and ‘Refiner’	46
3	Example prompt, applying concepts of ‘Stepwise Prompting’ . . .	47
4	Example prompt, applying concepts of ‘Zero-Shot’ Prompting . .	48
5	Example prompt, applying concepts of ‘Few-Shot’ Prompting . . .	49
6	Diagram of the Triangulation Strategy, from (Hills 2026)	77
7	Examples of effective prompt engineering techniques using in the stand-up comedy ‘standard template’ relating to the CLEAR and SMART frameworks.	88
8	Stand-up comedy generation workflow, categorized by AI cognitive processing modes.	89
9	This shows the typical structural flow of LLM-generated comedy routines.	100
10	Structural breakdown and prompt logic for the review generation process.	104
11	Diagram of the Triangulation Strategy, from “Stand-Up and Deliver” Lecture (Hills 2026)	121
12	The 25 distinct categories identified by (Yenidogan 2023).	123
13	A graph showing the average topic choice scores for model	124

14	A graph showing the scores for each topic choice across the 25 generations	125
15	A graph showing the rate of change across the “Self-Review / Self-Improvement” process per model	141
16	A graph showing the total change from the first to last iteration per model	141
17	A graph showing the Levenshtein Edit Rate across the iterations .	143
18	A graph showing the total change from the first to last iteration per model	143
19	A graph showing the Levenshtein Edit Rates for all the generations in the dataset	144
20	A graph showing the Semantic Distance across all the generations, with category zones	161
21	A graph showing the count of each category of Semantic Distance scores	161
22	A graph showing the the Semantic Distance scores of all individual jokes from the dataset, with zones	163
23	A graph showing number of annotations added per model	171
24	A graph showing number of normalised annotations added relative to routine length per model	172
25	A graph showing percentage of annotations added that weren’t specified in the prompt	176
26	Methodology of extracting jokes for the external survey	178
27	A graph showing the number of votes of each iteration, across wave 1180	
28	A graph showing the number of votes of each iteration, across wave 2181	
29	A graph showing the voting percentage of each iteration, across the two waves	182

30	A selection of qualitative comments with highlighting applied to map the distribution of positive, negative, and neutral sentiments.	189
----	---	-----

List of Prompts

4.1	A simple prompt for writing a poem about ‘Lavender’	80
4.2	A format-specified prompt for writing a poem about ‘Lavender’ .	81
4.3	A prompt for explaining the concept of ‘peer to peer networks’ . .	83
4.4	A prompt for explaining the concept of ‘peer to peer networks with intended audience context’	84
4.5	The initial prompt for generating stand-up comedy routines with LLMs	85
4.6	The template prompt for applying “persona” to stand up comedy generation	88
4.7	An initial prompt generating topics for stand-up comedy routines	90
4.8	A “persona”-driven prompt generating topics for stand-up comedy routines	91
4.9	Additional context for a prompt to include additional informative instructions	91
4.10	The final prompt for generating stand-up comedy topics	92
4.11	The initial prompt for generating stand-up comedy routines	95
4.12	A further developed prompt generating stand-up comedy routines	96
4.13	The final prompt for generating stand-up comedy routines	96
4.14	An alternative “relationship” variant of the prompt for generating stand-up comedy routines	97

4.15	An alternative ‘more information’ variant of the prompt for generating stand-up comedy routines	101
4.16	The final prompt for “reviewing” and “justifying” the review of a stand-up comedy routine	105
4.17	The final prompt for “refining” the draft of a stand-up comedy routine	106
4.18	The prompt used to create a “worse” redraft of the routine	110
4.19	The template prompt used to add annotations to the routine . . .	113
5.1	An extract of the prompt used to generate stand-up comedy material in Phase 2 of the experiment	130
6.1	An alternative prompt for reviewing the stand-up comedy routines using a more consistent approach	215
6.2	A prompt to simulate an audience member “heckle” and the LLM’s response	233

List of LLM Outputs

1	Basic “Lavender Field” Poem	80
2	Lavender Field Poem with Formatting	81
3	“Peer-to-Peer Networks” Definition	83
4	“Peer-to-Peer Network” Definition (Simplified)	84
5	Self-Deprecating Persona Example	86
6	Hyperble Example in Routine	86
7	“Closer” Example in Routine	87
8	“Aging” Premise Justification	93
9	“Inner Voice” Premise Justification	94
10	Bard Persona / Character Example	97
11	“Social Media Addiction” Extract	98
12	“Family Dating Pressures” Extract	99
13	“Modern Technology” and Annotations Extract	102
14	Summary of “Dating Apps” Routine Iterations	107
15	Tracking the “Amy” character across iterations	108
16	Annotated “Dating” Routine Extract	114
17	Annotated Stand-Up Comedy Routine Extract	115
18	Example comedic premise of “Aging”	125
19	Alternative Example comedic premise of “Aging”	125
20	Example comedic premise of “Social Media Overload”	126
21	Alternative Example comedic premise of “Social Media”	126

22	Alternative Example comedic premise of “Social Media Addiction”	126
23	Specific comedic premise of “Conspiracy Theories My Parents Believe In”	127
24	Specific Example comedic premise of “The Hunger Games of the Remote Control”	127
25	Three Examples of highly specific comedic premises	127
26	“Ice Cream Callback Closer” Extract	132
27	“Wi-Fi Callback Closer” Extract	132
28	“Rule of Three Romance + Technology” Extract	133
29	“Self-Deprecation Embarrassment” Extract	133
30	“Self-Deprecation Dating Apps” Extract	134
31	“Self-Deprecation Future” Extract	134
32	“Self-Deprecation Fitness” Extract	135
33	“Self-Checkout / Bermuda Triangle Incongruity” Extract	136
34	“Customer Service / Russian Roulette” Extract	136
35	“Dating Apps / E-Commerce” Extract	137
36	“Corporate / Summer Camp” Extract	137
37	“Tax Forms / Ancient Conspiracies” Extract	137
38	Example of “Closer” Improvement	146
39	Example of “Increasing Audience Interaction” Improvement	146
40	Example of “Refining the Routine’s “Punchline””	147
41	Example of “Adding Gym Personal Anecdotes”	147
42	Example of “Tightening Routine Deliveries”	148
43	Example of “Introducing Physical Comedy Elements”	148
44	Example of “Utilising Dynamic Act-Outs”	149
45	Example of “Adding Joke “Tags””	150
46	Example of “Adding Direct Crowd Work”	150
47	Example of “Adding a Narrative Callback”	151

48	Example of “Implementing Narrative Plot Twists”	152
49	Example of Semantic Distance in “Too Predictable” Zone	154
50	Example of Semantic Distance in “Light Obervation” Zone	155
51	Example of Semantic Distance in “Sweet Spot” Zone	156
52	Example of Semantic Distance in “Hallucination” Zone	157
53	Extract showing Annotations added that were not in the sample .	175
54	Example of ChatGPT writing a response to a Heckler	233

Chapter 1

Introduction

With the growing rise of humour-bots, the question of artificial intelligence attempting stand-up comedy is less explored than one may think. The questions explored throughout this thesis align with the rise in public interest in Generative Pre-trained Transformers (GPTs) since OpenAI’s release of GPT-3 (Brown et al. 2020). In the Computational Creativity (CC) field of research, we have seen LLMs tackle many creative endeavours such as fiction writing (Shakeri, Neustaedter and DiPaola 2021), poetry (Sawicki et al. 2022) and short-form joke generation, or Wit (Toplyn 2021), with varying degrees of success.

Most recently, the establishment of the Computational Humour (CHum) Workshop at ACL (Hempelmann et al. 2025) has signalled a distinct change in maturity of this field, officially recognising that computational humour is now a research-rich area of computer science, linguistics, deep learning and social ethics.

Despite these advancements, the stand-up comedy domain has largely remained the elusive domain of human comedians, untouched by LLM-prompts. As a medium of spoken word art, stand-up comedy has developed a unique playbook of techniques and styles. When considering the genre, it is clear that stand-up comedy is composed of far more than just “telling jokes”.

1.1 Bringing Computer Science and Stand-Up Comedy together

As a performer and computer scientist myself, the idea of combining stand-up comedy and generative models proved to be a novel and thought-provoking concept. Although I had experimented with AI-generated stand-up comedy material in the stand-up world from the angle of a performer, a more academic investigation was an appealing new perspective

When considering whether LLMs can truly recreate human creativity, it seems only fitting for stand-up comedy to be an area of interest for researchers in the field. Comedy invokes many different human traits to create an optimal product, and computational systems have mostly challenged the area of short-form comedy, rather than long-form stand-up. This is very apparent in the way that comedians write, review and improve their material before taking it on tour and during the tour.

However, this process remains somewhat unexplored with LLMs. An area of research, Computational Creativity, looks to explore the capabilities of these generative models, and little research has been done to give the community an understanding over how these models understand strengths and flaws in their writing, and how they may improve their generations.

While standard Large Language Model (LLM) generation can be viewed as a continuous flow of “Engagement”, it typically lacks the “Reflection” state necessary for human-based creative writing. This thesis aims to bridge that gap by simulating the Engagement-Reflection cognitive model, treating the LLM not merely as a static “joke machine,” but as an active curator of its own material. By automating a feedback loop of drafting, failing, analysing, and rewriting, we can observe whether these models are capable of demonstrating self-review and self-improvement in what they generate.

1.2 Research Questions

This somewhat unexplored area will form the key question of this thesis: **using LLMs that are publicly accessible, are they able to evidence self-review and self-improvement, through the domain of stand-up comedy?** These models will be investigated using a prompt-chaining pipeline inspired by the stand-up comedy writing process. Their outputs will be explicitly assessed to identify evidence of Jordanous and Keller’s “Progression and Development” component of creativity, which focuses on the advancement and evolution of an artifact during a process rather than just the final product. Because humour evaluation is inherently subjective, the success of this self-improvement will be measured using a triangulation strategy: evaluating structural validity against humour theories, quantitatively measuring the semantic gap, and longitudinally tracking the edit rates across iterations.

The specific idea of “Progression and Development” will focus on the way the outputs develop over the course of the process. It diverges from other existing research in the computational creativity field, which primarily focuses on joke generation, wordplay, and humorous chat-bots. Rather, this research concentrates on the **development process** of short stand-up comedy routines.

To develop these routines, four core elements are used:

- **Topic Generation**
- **Material Writing**
- **Iterative Improvement**
- **Performance Annotations**

The methodology of this thesis will explore these elements in more detail.

1.3 Thesis Structure

This thesis consists of seven chapters. Following this introduction, Chapters 2 and 3 will discuss the surrounding background and literature in this area while providing a passing overview of stand-up comedy and its traits. Chapter 4 will then dive into the methodology, creating a process by which to prompt LLMs to achieve the desired results. This methodology will be tested and evaluated in Chapter 5, before discussing the successes, limitations, and ethical implications of this research to the field in Chapter 6, as well as the scope for future research. Finally, Chapter 7 will conclude the thesis by bringing together the findings and drawing a final conclusion on the generative capabilities of large language models in the domain of stand-up comedy.

Before exploring the generation of computationally-generated stand-up material, we need to understand what the genre of stand-up comedy is. The next chapter will delve into this domain, exploring the history of the art form, the foundational theories of why we laugh, and the intricate mechanics used to construct and evaluate jokes. By establishing this theoretical groundwork, we can better understand the complex linguistic and performative rules these models are tasked with engaging with.

Chapter 2

Background

Before exploring the creation of computationally-generated stand-up material, we need to understand what the genre of stand-up comedy is. This chapter discusses the domain of stand-up comedy and the work surrounding it. The goal of this chapter is to illustrate the history of stand-up comedy, what we look for in comedy routines, and how we evaluate jokes.

2.1 The Art and History of Stand-Up

2.1.1 A History of Stand-Up Comedy

The concept of a ‘comedian’ exists far outside the stand-up comedy landscape, including comedy sketch artists, such as BYU TV’s Studio C (BYUtv 2012), to late-night talk show legends like Conan O’Brien (O’Brien 2010). However, the research in this thesis primarily focuses on comedy within the domain of the stand-up comedy genre and the role of a stand-up comedian. Stand-up academic Oliver Double has dated the term of “stand-up comedian” to as early as 1917, in an article written by The Yorkshire Evening Post (Double 2018).

While the specific terminology is relatively modern, the roots of the practice have been traced back to the 15th-century Italian *commedia dell’arte*, later evolving

through British music hall traditions and American Vaudeville (Schwarz 2010). At the start, these performances were embedded within larger variety acts that included music and burlesque. However, later on, stand-up comedy developed as a distinct genre by isolating the spoken parts from these theatrical variety shows. They stripped away the costumes and dramatic vehicles to leave the performer exposed (Schwarz 2010).

This development cemented the idea of the solo performer: an individual who plants themselves in front of an audience to deliver a monologue of stories or jokes. This role underwent a transformation in the mid-20th century. Sociological analysis suggests that the modern stand-up we are familiar with today began with figures like Lenny Bruce, who shifted the genre from mere joke-telling to a platform where individuals “speak on their own behalf” (Smith 2018). This shift established the comedian not just as an entertainer, but as a social agent who uses that platform to turn social identity and power into individual habits, quirks, traits and characteristics, (Smith 2018).

2.1.2 Stand-Up as a Linguistic Genre

While performance factors like timing and physicality play a role, modern stand-up comedy has been defined as a “linguistic genre” (Schwarz 2010). Scholars argue that stand-up is essentially “a form of talk” where humour is constructed through “linguistic and rhetorical structures” rather than physical slapstick (Wawrzyniuk 2021).

Despite the illusion of spontaneous conversation, stand-up is characterised as a “highly artificial, scripted genre” (Schwarz 2010). The humour relies on “linguistic features of joke telling,” such as wordplay, ambiguity, hyperbole, and formulaic repetition (Schwarz 2010). Because the core of a stand-up routine (the script) is built upon these textual patterns and verbal structures, this domain is suitable for the application of Large Language Models (LLMs). Since LLMs excel at processing

and generating complex text sequences based on learned linguistic probabilities, they may be well-positioned to replicate the “highly scripted” humorous narratives that underpin the art form (Wawrzyniuk 2021).

2.1.3 The Persona

On the question of what ‘Stand-Up Comedy’ means, many experts have varying definitions. Academic Oliver Double defined stand-up comedy with three key elements: “Personality, Direct Communication, and Present Tense” (Double 2021). This pertains to the fact that a stand-up comedian should have a clear stage ‘persona’, have direct communication with the audience, and the performance should be framed from the perspective of the current moment. This suggests that a good direction for the prompt engineering aspects of this research will be to consider how a ‘persona’ can be brought out of the LLM.

The balance of a stage persona is often framed as a tension between authenticity and character. While historically, comedy figures were defined by fixed archetypes like the “clown” or the “fool”, modern stand-up has shifted toward an “intra-personal” art form where the performer speaks on their own behalf (Smith 2018). Unlike theatrical acting, where the role is enacted, the stand-up persona often removes the boundary between art and reality, presenting a “heightened version of self” rather than a mask (Smith 2018). This spectrum ranges from comedians like Jerry Seinfeld, whose stage persona closely mirrors his off-stage personality, to those like Steven Wright, who adopts a detached, ‘deadpan’ character to deliver fabricated, surreal anecdotes (Schwarz 2010).

The distinction between authenticity and character is directly relevant to the application of large language models in comedy generation. As effective prompting requires defining who we want the AI to ‘pretend’ to be, the prompt must act as a directorial constraint that establishes the model’s simulated identity. This will be discussed further in Chapter 4, when prompting is discussed in more detail.

Since stand-up comedy is fundamentally an art of “intra-personal” relations where a “self-other” dynamic is built, an LLM cannot simply generate text in a vacuum. It must simulate a ‘self’ to relate to the audience. Just as a human comedian develops a specific ‘stage persona’ to organise and perform their material. Whether simulating a ‘deadpan’ delivery or an ‘observational’ style, the prompt must supply the specific “biographical suspense” or character constraints that allow the model to replicate the comedic aesthetic of a modern stand-up comedian.

2.1.4 Styles of Stand-Up

There are different styles of stand-up comedy, all subverting the audience’s expectations in different ways. The common types of stand-up predominantly fall into the following genres:

- **One-Liners:** Engage a shorter style of joke that suits more abrupt comedians, such as Jimmy Carr. These rely on fast paced incongruity to deliver a punchline.
- **Observational Comedy:** Favoured by comedians such as Michael McIntyre, this is a genre that explores the “everyday phenomenon that is rarely discussed” (Double 2021).
- **Anecdotal Comedy (Storytelling):** This refers to comic personal stories that may be true or partly true but embellished. Unlike the rapid-fire structure of one-liners, storytelling routines are “more complex and elaborated” utilising a structure similar to “folk talk” or popular storytelling (Wawrzyniuk 2021). These narratives often include an orientation, a complication, and a resolution, requiring a longer setup to establish the context before delivering the punchline.
- **Surreal or Absurd:** This style relies on “totally bizarre and awkward content”, often presenting “fake life situations” that the comedian has never

actually experienced (Schwarz 2010). In this style, the humour is derived not from relatability, but from the “incongruity expressed in the nonsense”, leaving the audience with a sense of absurdity rather than a reflection of reality (Schwarz 2010).

- **Deadpan:** Often overlapping with the surreal, the ‘deadpan’ style is characterised by an “apathetic”, “monotonous”, and “expressionless” delivery where the comedian conveys no warmth or connection to the audience (Schwarz 2010). This style uses a lack of emotion and physical movement to heighten the contrast between the mundane delivery and the often ridiculous content of the jokes.

Because stand-up comedy encompasses such distinct sub-genres, evaluating the output of an Artificial Intelligence model requires a clear understanding of the specific ‘style’ it is attempting to generate. You cannot judge an unusual joke by the metrics of a one-liner, as they operate on different fully linguistic and structural principles. Similarly, a ‘deadpan’ or ‘surreal’ routine might intentionally utilise totally bizarre and awkward content that fails standard coherence metrics but succeeds through absurdity. Therefore, any computational evaluation framework must be aligned with the intended genre to accurately measure success, rather than applying a universal metric that might penalise a valid stylistic choice, such as a long pause or a nonsensical premise, as an error.

This section paints a picture of the stand-up comedy domain that my research will use as a foundation to assess the capabilities of large-language models. Understanding what the ‘comedian’ is informs the different traits and characteristics to look for within later prompting and LLM outputs. Looking at the different styles of comedy shows how this research fits alongside other research within this genre.

2.1.5 The Narrative Arc

Some of these styles rely upon a ‘Narrative Arc’ to ensure success in the setup-punchline structure, where this arc is forms of a comedian combining jokes of a similar topic or idea into longer form routines. These routines often tell a story and more heavily rely on anecdotes. Although favouring anecdotal and observational, most styles can make use of this Narrative Arc. Even within some one-liner acts, a comedian will group together the one-liners by topic or similarity to help provide structure to the routine. This applies to the distinction between one-liners (‘wit’) and anecdotal comedy, which will be discussed further in Chapter 3.

2.2 Foundational Humour Theories

2.2.1 Why do we laugh?

American comedian Lenny Bruce suggested that a comedian is “one who performs words or actions of his own original creation, usually before a group of people in a place of assembly, and these words or actions should cause the people assembled to laugh at a minimum of one laugh every 25 seconds for a period of not less than 45 minutes, and accomplish this feat with consistency 18 out of 20 shows” (Double 2021). This description can be found to be vastly over-defined and creates a set of parameters that doesn’t allow for the exciting variation in style and delivery of stand-up comedy. To better understand the mechanisms behind laughter, we look to academic theories.

The Incongruity-Resolution Theory

The Incongruity-Resolution Theory is one of the most influential framework for analysing stand-up comedy, as it addresses the cognitive process of the setup and punchline. At its core, this theory posits that humour arises from “incompatible

and contrastive ideas” or an “unfulfilled expectation” (Schwarz 2010). Historically, this concept traces back to philosophers like Kant, who argued that laughter is an affection arising from the sudden transformation of a strained expectation into nothing, and Schopenhauer, who focused on the mismatch between abstract concepts and real objects (Schwarz 2010).

When looking at stand-up comedy specifically, the theory is often shown as a two-stage cognitive process: first, the perception of incongruity, and second, the resolution of the setup.

The comedian’s strategy is to “build up a joke with the intention of leading the audience to a false expectation” (Schwarz 2010). When the punchline is revealed, it provides an “unexpected outcome,” confronting the audience with a surprise that resolves the tension created by the setup (Schwarz 2010). This ‘surprise’ / ‘unexpected’ quality will be interesting to explore in the context of LLMs being next-token generators, aiming for predictable outputs.

The General Theory of Verbal Humour (GTVH)

The Incongruity-Resolution framework is essential for understanding the General Theory of Verbal Humor (GTVH), developed by Attardo and Raskin in 1991 (Attardo and Raskin 1991). The GTVH represents a comprehensive linguistic framework for analysing comedy text. This theory is an expansion of previous work on the Semantic Script-based Theory of Humour (SSTH). The earlier model focused exclusively on the semantic opposition between scripts, while GTVH acknowledges that humour is a multi-faceted phenomenon that requires a broader set of parameters to be understood (Wawrzyniuk 2021).

The GTVH dictates that humour cannot occur without “Script Opposition” (SO). SO refers to the presence of two overlapping but opposing scripts (e.g., normal/abnormal, possible/impossible) compatible with the same text (Wawrzyniuk 2021). As Wawrzyniuk notes, “no incongruity means no amusement”. This means

the discrepancy between these scripts is the central engine of the joke, making this theory particularly relevant for analysing the linguistic structures used by LLMs to generate humour.

The GTVH states that any ‘funny’ text can be analysed according to six parameters, known as “knowledge resources” Attardo and Raskin (1991). These resources provide a formal vocabulary for deconstructing exactly why a joke works or fails.

1. **Script Opposition (SO)**

This refers to the central conflict or contrast within the joke.

2. **Logical Mechanism (LM)**

This explains how the two conflicting scripts in the SO are brought together, allowing the listener to resolve the incongruity.

3. **Situation (S)**

This is the characters, the setting, the objects, and the activities bring described within the joke.

4. **Target (T)**

This is the ‘butt’ of the joke, be it person, group, institution, or stereotype.

5. **Narrative Strategy (NS)**

This is the genre of format of the joke. Is it a story, or a pun, or a dialogue or a one-liner?

6. **Language (LA)**

This the verbal delivery of the joke, including the choice of words, syntax, phonology and placement of the punchline.

Scholars argue that these resources are hierarchical. The SO is essential and abstract, whereas the Language is variable. This means in practice that a

comedian can tell the same joke (same SO, LM and T) but using different words (LA). These parameters provide a qualitative foundation for analysing stand-up comedy material.

Benign-Violation Theory

A more recent but highly influential framework in contemporary humour studies is the Benign Violation Theory (BVT), developed by Peter McGraw and Caleb Warren. This theory combines elements of incongruity, superiority, and relief theories into a single model, suggesting that humour occurs when and only when three conditions are satisfied (McGraw and Warren 2010):

1. A situation is a violation.
2. The situation is benign.
3. Both perceptions occur simultaneously.

In this context, a “violation” refers to anything that threatens a person’s well-being or sense of how the world ‘ought’ to be. This ranges from physical threats (like tickling) to breaches of social norms (like a crude joke) or linguistic rules (like a pun). However, for these violations to be funny rather than purely offensive or frightening, they must be perceived as ‘benign’ (safe, acceptable, or inconsequential) (McGraw and Warren 2010). This spectrum is shown visually in Figure 1.

For stand-up comedy, this theory is particularly valuable in explaining the ‘ideal space’ that comedians must navigate. A joke that is too benign (such as a boring statement of fact) fails to provoke laughter because there is no threat or violation to engage the audience. On the flip side, a joke that is a pure violation (e.g., a statement that is genuinely hurtful, racist, or terrifying without a playful context) will result in negative affect rather than amusement (Renaissance 2024).

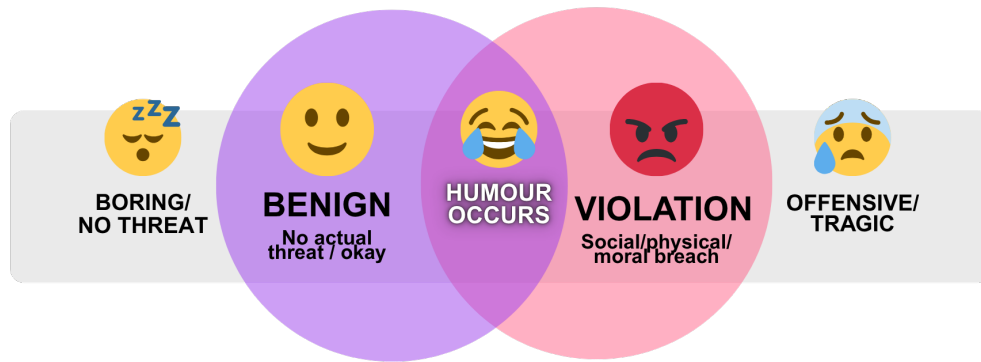


Figure 1: Diagram of the Benign-Violation Theory, from (Hills 2026)

Comedians often transition severe violations into acceptable jokes through specific techniques, such as psychological distance (e.g., “Comedy is tragedy plus time”), the context of the comedy club (a “play frame”), or by targeting low-commitment norms (McGraw and Warren 2010).

Critics of this theory argue that it can be overly broad or reductionist. For instance, while BVT explains why we laugh at a person slipping on a banana peel, it sometimes struggles to account for the nuances of complex narrative humour or the subjective nature of what an individual considers acceptable to joke about based on their personal or cultural background (Umam, Rusliyanti and Oktaviani 2025). However, despite this critique, BVT remains a robust tool for analysing why jokes might fail, because the comedian produces material that is either too safe (benign but no violation) or inadvertently offensive (violation but not benign).

Superiority Theory

While theories like the Incongruity-Resolution Theory focus on cognitive mismatch, Superiority Theory focuses on the social and emotional drivers of laughter. This framework suggests that humour arises from a sense of superiority derived from the misfortune, stupidity or failure of others (Schwarz 2010). The roots of this theory trace back to classical philosophers such as Plato and Aristotle, who argued that laughter is inherently malicious and born from a mixture of pleasure and pain when observing the shortcomings of others (Attardo 2020).

Thomas Hobbes speaks about laughter as a “sudden glory” arising from the conception of some superiority in ourselves by comparison with the weakness of others or with our previous selves. In the context of stand-up comedy, this theory explains the mechanics of aggressive humour, ridicule and ‘roasting’. When a comedian mocks a heckler, a political figure, or a social group, they are inviting the audience to join a temporary ‘in-group’ that feels superior to whoever the joke is targeting (Schwarz 2010).

This theory also extends to include self-deprecating humour, a frequently employed technique used by comedians. When a comedian mocks themselves, they are essentially splitting their ego into two parts. The speaking ‘self’ mocks the past ‘self’ based on an event that has usually already happened, allowing the audience to laugh from a position of superiority over the comedian’s (often manufactured) incompetence.

Relief Theory

Relief Theory approaches humour from a physiological and psychological perspective. This theory suggests that laughter functions as a mechanism to release excess nervous energy or tension (Schwarz 2010). Herbert Spencer argued that emotions produce physical energy that must be discharged. In humorous situations, where the threat is perceived as non-serious, the built-up energy is ‘vented’ through

laughter.

For the stand-up comedian, this theory relates to discussing ‘taboo’ subjects such as death, sex, or tragedy. The comedian deliberately builds tension in the room (the ‘setup’) and then provides a release (the ‘punchline’). The laughter generated is directly proportional to the amount of tension the comedian manages to build before the release (Cela, Veit and Wood 2024).

2.2.2 Summary

Each of these frameworks provides a specific lens through which to view comedy. Scholars increasingly argue that no single theory is sufficient to explain the complex nature of stand-up comedy. Instead, stand-up comedy should be viewed with a combination of these theories rather than as competing definitions. A single joke may rely on the Incongruity-Resolution Theory for its structure, while simultaneously drawing on the Superiority Theory to define its social target and Relief Theory to manage the tension of the delivery. These provide evidenced frameworks for evaluating the techniques used in stand-up comedy material. This will inform the qualitative analysis discussed further in the thesis, as a way of measuring stand-up comedy.

2.3 The Mechanics of the Joke

2.3.1 The Structures of a Joke

Stand-up routines typically consist of a ‘premise’, an application of techniques, and a ‘punchline’. The function of the premise is to narrow down the topic into more specific subjects (Dean 2024). Examples include a recent breakup (Graham 2024) or getting called for jury service (Acaster 2018).

The punchline is delivered after a setup, which intends to mislead an audience

into a false assumption, which the punchline will then disrupt, leading to laughs. Following a punchline, a ‘tag’ can be used to achieve a second punchline without needing any additional setup (Dean 2024).

The Premise

The premise is the underlying foundation upon which the expectation rests. In the practical guide *Comedy Terms*, Greg Dean distinguishes the premise from the broader topic of a routine. While a topic might be general (e.g. relationships or air travel), the premise is defined as “a narrower idea stated as a negative opinion about a subject” (Dean 2024). This “negative opinion” is crucial because it generates the initial tension or conflict required for a narrative arc. It establishes a specific point of view that the comedian intends to explore or defend.

The premise serves to establish the expected mode of communication before the switch to the humorous mode occurs. It acts as the anchor for the first script from the General Theory of Verbal Humour, representing the normal / expected reality that the audience needs to accept. For a joke to succeed, the audience needs to buy into their initial reality (“buy into the premise”) regardless of whether it is a truthful observation or surreal logic. If the premise is not accepted or understood, the subsequent punchline has no baseline to contrast against. The premise is not merely the introduction to a joke, but rather the contract of reality established between performer and audience.

The Callback

The ‘callback’ is widely considered one of the most powerful tools in a comedian’s toolbox, serving as a structural device that transforms a collection of isolated jokes into a cohesive routine. It is defined as a joke that references a piece of material delivered earlier in the set (Dean 2024). However, its function goes beyond mere repetition, acting as a second punchline that requires no setup, similarly to a tag.

The advantage being that this second punchline can occur later in the routine rather than directly after the previous punchline. The reason for this is that the context of the previous joke has already been established and resides in the audience's short-term memory.

The callback is essential for unifying a performance. While novice writing often results in disjointed lists of jokes, the use of callbacks can help create a sense of narrative closure. As noted in practical guides, saving a strong callback for the end of the show is a “safe strategy” to ensure the performance ends on a high note, making the entire set feel more cohesive and considered (Side Splitters Comedy 2024).

Academically, the callback functions through the mechanism of intertextuality and “lexical recurrence” (Schwarz 2010). By repeating a specific phrase or concept in a new context, the comedian triggers a memory recall in the audience. On a connection level, this process creates a unique form of ‘in-group’ similarly to Superiority Theory. The audience laughs not only at the punchline but at the shared recognition of the reference.

2.3.2 Linguistic Techniques

To transform a premise into a successful joke, comedians rely on specific linguistic strategies that manipulate language to create ambiguity, misdirection, and surprise. These techniques function as the “Language” (LA) knowledge resource within the GTVH framework, acting as the mechanism that delivers the Script Opposition (Attardo and Raskin 1991).

Wordplay and Puns

Wordplay serves as the most fundamental form of linguistic humour, relying on the manipulation of the phonetic or semantic properties of words. Primarily, wordplay exploits “lexical ambiguity”. This refers to the fact that a single word

can have multiple meanings (polysemy) or that different words can sound identical (homophony) (Toplyn 2021).

The ‘pun’ is the most distinct example of this technique. A pun utilises a “connector” (a word or phrase that functions as a bridge between two incompatible scripts) (Schwarz 2010) to achieve its humour. For a pun to be successful, the audience must initially interpret the connector in one context (Script A), only to be forced by the punchline to reinterpret it in a second context (Script B). While often dismissed as a “lower” form of humour, puns are significant to computationally generated humour because they are among the easiest forms of humour for AI to generate. As demonstrated by the Witscript system, algorithms can effectively identify keywords with multiple definitions and construct sentences that link them, making wordplay a primary domain for early computational humour research (Toplyn 2022).

Double Entendres

The double entendre is a figure of speech where a spoken phrase can be understood in two different ways, with one meaning usually being transgressive, risqué, or ironic. Within the GTVH, this relies heavily on “Script Opposition” to trigger a conflict typically between a “Socially Acceptable” script and a “Sexual/Taboo” script (Schwarz 2010; Attardo and Raskin 1991). The mechanism works by exploiting specific lexical triggers. The comedian constructs a sentence that makes grammatical sense within the innocent context (the setup) but uses specific words that activate the taboo association in the audience’s mind.

Hyperbole

Hyperbole, or extreme exaggeration, is a staple of anecdotal stand-up comedy. It involves the “magnification” of a situation or characteristic to an absurd degree to highlight a truth (Schwarz 2010). Unlike lying, which intends to deceive, hyperbole

intends to emphasise. It signals to the audience that they are entering a “play frame,” allowing the comedian to explore the emotional reality of a situation rather than the factual reality. Linguistically, this is often achieved through the use of “extreme case formulations” (e.g., “always,” “never,” “everyone”) and quantitative exaggeration (Wawrzyniuk 2021). By blowing a minor annoyance out of proportion, the comedian creates an incongruity between the triviality of the event and the intensity of the reaction.

Implication and Paralipsis

Implication refers to the strategy of leaving the most critical part of the joke unsaid, forcing the audience to “fill in the blanks.” This relies on the “principle of economy” in language; the humour arises from the cognitive effort the audience must expend to decode the meaning (Schwarz 2010). This technique furthers the strong bond between performer and audience / ‘in-group’ intimacy. If the comedian explicitly states the punchline, the joke may fall flat or come across as patronising. By using implication, the comedian trusts the audience’s intelligence to bridge the gap between the setup and the conclusion. This often manifests in “trailing off” sentences or leaving the final word of a sentence unspoken, relying on shared cultural knowledge to complete the thought (Schwarz 2010).

Paralipsis (also known as “apophasis”) is a type of implication where the speaker emphasises a subject by claiming to pass over it or denying that they are mentioning it (e.g., “I won’t mention my opponent’s corruption”). In stand-up, this functions as a form of “pragmatic insincerity” (Schwarz 2010). This technique allows the comedian to introduce taboo or insulting material while maintaining a facade of politeness or detachment. The literal meaning of the sentence (the denial) clashes with the speaker’s communicative intent (the accusation), forcing the audience to decode the implication and the unsaid meaning lying beneath the surface text.

Rule of Three

The Rule of Three (or comic triple) is arguably one of the most well-known structural device in Western stand-up. It exploits the cognitive mechanism of pattern recognition: the first element starts a pattern, the second confirms it, and the third disrupts it (Schwarz 2010). This structure serves as a condensed version of the setup-punchline arc. The first two items establish the expected script, while the third item functions as the incongruous element that shatters this expectation. As noted in research around computational generation, the brevity of this structure makes it a prime candidate for template-based generation, yet the challenge lies in ensuring the third item provides a sufficient semantic shift to trigger the ‘surprise’ necessary for laughter (Winters, Nys and De Schreye 2018).

Simulated Spontaneity

A crucial linguistic aspect of stand-up is the illusion of spontaneity. Although routines are highly scripted, comedians use Reformulation (correcting oneself mid-sentence) to make the monologue appear to be occurring in the present tense (Schwarz 2010). Phrases like “I mean,” “or rather,” or “what’s the word?” are inserted deliberately to mimic natural conversation. This “pseudo-spontaneity” lowers the audience’s critical defences, making the eventual punchline feel like a shared discovery rather than a rehearsed recitation (Smith 2018).

2.3.3 Summary

Collectively, these linguistic techniques demonstrate that stand-up comedy is a highly structured ‘linguistic game’, defined by key principles and patterns. For the purposes of this thesis, these techniques can be checked for and identified within AI-generated comedy as with human-generated comedy, meaning that this can be developed into a measure of how successful a stand-up script is. Identifying these

techniques within LLM-generated stand-up comedy material will be crucial in concluding its success or failure in generating material in the domain of stand-up comedy.

2.4 The Performance Layer

2.4.1 The Text-Performance Gap

American Academic Lawrence E Mintz defined stand-up comedy as an “encounter between a single, standing performer behaving comically and/or saying funny things directly to an audience, unsupported by very much in the way of costume, prop, setting, or dramatic vehicle” (Double 2021). This is considered to be a rather outdated definition but acknowledges many successful elements of the genre. In recent years we have seen people use a variety of props and costumes throughout routines, such as the recent rise in musical and PowerPoint comedy, the latter where my own comedic adventures fall.

While this thesis primarily focuses on the linguistic capabilities of Large Language Models (LLMs), it is critical to acknowledge the fundamental “text-performance gap”, important in the genre of stand-up comedy. Unlike written humour found in essays or novels, where the text is a self-contained entity designed to be consumed privately by a reader, a stand-up routine is an incomplete artifact until it is performed (Schwarz 2010).

Academic analysis suggests that a stand-up script functions less like a novel and more like a “musical score” or a “skeleton” script (Smith 2018). The written words provide the structural logic (the set-up, the punchline, and the sequencing) but they lack the paralinguistic information required to fully realise the humour. Elements such as intonation, timing, facial expressions, and the ‘comic pause’ are not explicitly encoded in the text but are superimposed by the performer during the act (Schwarz 2010).

This distinction creates a specific challenge for computational generation. A novel is “written to be read,” whereas stand-up comedy is “written to be spoken” (Schwarz 2010). Consequently, when an LLM generates a routine, it is creating a “decontextualised” text. It lacks the “here and now” sociality: the specific relationship between the performer and the immediate audience that defines the live experience (Smith 2018). Therefore, identifying a generated text as stand-up comedy requires the reader to mentally perform the text, filling in the missing gaps that a human comedian would naturally supply.

2.4.2 Writing vs. Performing

The next stage in crafting a stand-up routine, is taking the script from the written page to the live stage. While a script may contain the necessary semantic logic (the setup and the punchline), it has been described as an inert “skeleton” until it is animated by performance (Smith 2018). In literary humour, such as a satirical novel, the reader controls the pace of consumption. In stand-up, the performer dictates the rhythm, meaning that a joke which reads perfectly on paper may fail disastrously in a club if the “oral punctuation” is mismanaged (Schwarz 2010). This distinction creates a division between the “authorial self” (who writes the material in isolation) and the “performing self” (who delivers it in front of a crowd). This translation from written words to performance goes beyond the scope of this thesis and will be discussed further within the Future Work discussion in Chapter 6.

2.4.3 The Autonomy of the Written Script

However, it is essential to distinguish between the ‘performative incompleteness’ of a script and its inherent merit as a literary and structural artifact. While the performance layer is the traditional vehicle for delivery, the work of the comedy writer possesses a formal autonomy that exists independent of the eventual live

execution. This ‘writerly merit’ resides in the semantic architecture of the joke, specifically the precise linguistic engineering required to establish expectations and trigger the resolution of incongruity. In the professional industry, this independence is shown by the existence of staff writers and “punch-up” specialists who generate value through the written word alone, often never intending to perform the material themselves. For the purposes of evaluating Large Language Models, this suggests that a comedic text can be analysed as a successful ‘skeleton script’ of humor. If the generated text contains the necessary logical markers, rhythmic cues, and structural integrity to be recognized as comedy by a reader, it holds status as a valid comedic artifact, even when stripped of its sonic and physical dimensions.

2.4.4 Paralinguistics

While the audience interaction element is beyond the scope of this project and will be discussed in future work. It is however important that we understand the key paralinguistic features of stand-up comedy, especially for a later stage in the LLM generations.

Timing and the ‘Comic Pause’

While the text provides the semantic content of a joke, the ‘comic pause’ is the mechanism that controls its cognitive processing. A pause is not merely an absence of sound but a deliberate “communicative act” that serves a dual function (Schwarz 2010). Firstly, it acts as a “processing window,” granting the audience the necessary milliseconds to resolve the incongruity presented in the punchline. If the comedian speaks too soon (“stepping on the laugh”), they disrupt this cognitive resolution, causing the joke to fail. Secondly, the pause builds psychological tension.

Drawing on Relief Theory, the silence before a punchline acts as a pressure cooker; the longer the silence (up to a critical point), the greater the tension and the subsequent explosive release of laughter (Schwarz 2010).

Intonation and Emphasis

Intonation functions as the “melody” of the routine, marking the structural boundaries of the joke. Comedians utilise specific “pitch contours” to signal to the audience where the setup ends and the punchline begins (Schwarz 2010). Typically, a setup is delivered with a rising or sustained pitch (indicating more is to come), while the punchline is delivered with a falling pitch and increased volume, signaling finality and the cue to laugh. This “oral punctuation” is critical for comedians, as a text lacking these markers may be semantically funny but performatively flat (Smith 2018).

Rhythm

Rhythm in stand-up comedy refers to the “pulsation” of the performance, otherwise known as the interplay between the comedian’s speech rate and the audience’s laughter (Schwarz 2010). A successful routine establishes a metronomic beat, often described by comedians as “getting into a roll.” This rhythm is collaborative; the comedian must adjust their delivery speed in real-time to match the “laughter decay” of the audience, ensuring they begin the next setup exactly as the previous laughter fades (Smith 2018). This dynamic synchronisation is what leads scholars to describe stand-up scripts as “musical scores” rather than static texts.

Physicality & “The Act-Out”

The ‘Act-Out’ is a specific performance technique where the comedian temporarily abandons the role of the narrator to “embody” a character or object within the story. Linguistically, this is defined as where the “I” of the speaker shifts from the comedian to the character (Schwarz 2010). During an ‘Act-Out’, the comedian utilises signs, such as changes in posture, facial expression, and voice quality, to enact the scenario rather than describe it. This shifts the mode of communication from “diegesis” (telling) to “mimesis” (showing). For example, instead of saying

“my father was angry,” the comedian adopts the father’s growling voice and rigid posture. This technique is vital for anecdotal comedy as it heightens the realism and immediacy of the story, allowing the audience to witness the event rather than just hear about it (Schwarz 2010).

2.4.5 Summary

These paralinguistics will naturally occur within generated stand-up comedy material, such as an LLM creating a story that involves acting out more than one character. It is important that we understand these techniques for the purpose of evaluation but understand that they will be disconnected from their ‘performing’ purpose.

2.5 The Process of Refinement

2.5.1 The Feedback Loop

Unlike other artists who may perfect their work in private before a public unveiling, the stand-up comedian is generally forced to practice in public or workshops. This process occurs primarily in ‘open mic’ nights, which function effectively as laboratories for data collection. In this environment, the audience acts as a collective editor, providing immediate, binary feedback: laughter or silence (Smith 2018).

This feedback loop is brutal but efficient. It serves as a “validity check” for the comedian’s premises. If a comedian presents a ‘benign Violation’ but the audience stays silent, the comedian knows immediately that the violation was either too benign (boring) or too severe (offensive). This “audience-as-editor” dynamic means that a stand-up routine is never truly a solo creation. Rather it is a “co-authored” text where the audience’s reactions dictate the final shape of the material (Schwarz

2010). The open mic allows the comedian to calibrate their ‘internal auditor’, training their own sense of humour against the reality of the public consensus.

This refinement process can be formalised into an iterative cycle:

1. **Generate**

The comedian writes a new premise and punchline based on an observation.

2. **Perform**

The material is tested live. This is the “stress test” where the linguistic techniques (puns, paralipsis, etc.) are deployed.

3. **Fail**

Failure is an expected part of the process. “Dying” on stage occurs when the Incongruity-Resolution process fails to trigger.

4. **Analyse**

The comedian must diagnose why the joke failed. They must determine if the failure was in the Script Opposition (the idea wasn’t clear), the Language (the wording was clumsy), or the Performance (the timing was off) (Attardo 2020).

5. **Rewrite**

The comedian adjusts the variable (i.e. by changing a word, shortening the setup, or altering the intonation) and prepares to re-enter the cycle at step 2.

This cycle highlights that comedy is a dynamic process of ‘trial and error’, where the final routine is the result of hundreds of micro-failures (Side Splitters Comedy 2024). This is the feedback loop that will be emulated through prompting in the methodology, with a few necessary changes / adaptations to fit a written-text-only approach.

2.5.2 ‘Tightening the Set’

The final stage of this process is known as ‘tightening the set’. Once a joke works, the goal shifts from effectiveness to efficiency. This involves the linguistic principle of “economy,” where the comedian removes any word or pause that does not directly contribute to the laugh (Schwarz 2010).

The objective is to maximise “Laughs Per Minute” (LPM). Comedians will strip away “connective tissue” (conversational filler used to link topics) to bring the punchlines closer together (Dean 2024). A “tight” five-minute set contains no wasted breath; every sentence is either a setup or a punchline. This optimisation process transforms a loose collection of anecdotes into a dense, high-impact performance.

For the purpose of this thesis, ‘tightening’ represents one of the biggest challenges for LLMs: it is not enough to generate text that is funny; the model must eventually learn to edit that text, simulating the human professional’s dedication to the concise delivery of the punchline and producing the maximum “Laughs Per Minute” ratio. (Side Splitters Comedy 2024).

2.6 Chapter Summary

This chapter has established that stand-up comedy is not merely the recitation of funny words, but a complex interaction between a ‘linguistic genre’ and a ‘performance art’. The literature identifies that a stand-up routine is an unstable text. It begins as a skeleton script, constructed using specific linguistic knowledge resources (like the pun, the rule of three, or paralipsis) to create an incongruity. However, the success of this script is partially dependent on the ‘performance layer’ (the timing, intonation, and persona) applied by the comedian in the moment.

This highlights that the creation of comedy is an iterative, data-driven process. Through a ‘feedback loop’, the audience typically acts as a co-author, validating or

rejecting the comedian’s premises. The transition from a raw idea to a ‘tightened set’ requires the comedian to survive failure, analyse the silence, and refine the economy of the language.

This provides the groundwork idea for this thesis: the journey, the refinement and how the material gets built. To keep the scope of this thesis manageable, this work will stop prior to the ‘performance layer’ but will explore the process up to that point, of writing and refining stand-up comedy material. The next chapter will explore how this area of research (humour, comedy and performance) can be brought into the domain of LLMs and Computational Creativity.

Chapter 3

Literature Review

When tackling this question of exploring whether LLMs can review and improve, it is crucial to understand what work has been done in the area of creating comedy using computational models and systems. This chapter will give a detailed summary of research and experimentation done in computer-generated humour, including artists, academics, and other types of performers. It will also discuss research that has been done in the evaluation of computational systems with creative outputs, as well as the ethical context of the work in this thesis.

3.1 Computational Creativity Frameworks

3.1.1 Defining Computational Creativity

Attributes of Creativity

In 2007, Ritchie developed a set of empirical criteria for evaluating creativity, which focuses on the observable properties of the system’s output rather than its internal processing (Ritchie 2007). Ritchie argues that for a program to be considered creative, it must first demonstrate:

- ‘typicality’ : the ability to produce items that are recognisable instances of

the target class

- ‘quality’ : the value of the item being produced
- ‘novelty’ : dissimilarity to existing examples used to train the program

Ritchie decomposes novelty into “untypicality” (deviation from the norms) and “innovation” (producing items not present in the inspiring set used to build the program) (Ritchie 2007). However, a limitation of this framework in the context of LLMs is the vastness of the training data. Proving innovation (that the item was not in the training data) is computationally intractable (cannot be solved efficiently by any known algorithm) for models trained on terabytes of internet text.

In Ventura’s paper on “mere generation”, he built on this existing research by outlining a definition of creativity of “Novelty, Value and Intentionality” (Ventura 2016). Ventura critiques systems that rely on ‘mere generation’ (randomisation or simple memorisation) and proposes that a truly creative system must satisfy three conditions. It must produce artifacts that are:

- new or original (Novelty)
- useful or meaningful within their domain (Value)
- generated through a deliberate process where the system has a goal or objective (Intentionality)

This suggests that a system is creative if the result is new/original, has merit within the domain, and the output has been generated deliberately.

This definition is tricky to explore when applied to large language models. While an LLM can produce results showing novelty and value, intentionality is difficult to attribute to the probabilistic engine behind the LLM models. However through the Witscript system, Toplyn argues that the system demonstrates creativity because

it intentionally filters outputs to ensure they meet a threshold of humour (value) and generates responses unique to the conversation (novelty) (Toplyn 2021). This filtering process simulates intent by rejecting the non-jokes. However, this could be argued as ‘curation’ by the programmer rather than intent from the agent. This demonstrates the subjective conflicting nature of evaluating creativity in a computational system.

Jordanous & Keller’s Components of Creativity

According to Jordanous and Keller, there are 14 key themes and factors of creativity identified through a corpus-based analysis of academic papers on the subject (Jordanous and Keller 2016). These components provide a “common language” for discussing creativity across different domains. The 14 components are:

1. Active Involvement and Persistence
2. Dealing with Uncertainty
3. Domain Competence
4. General Intellect
5. Generation of Results
6. Independence and Freedom
7. Intention and Emotional Involvement
8. Originality
9. Progression and Development
10. Social Interaction and Communication
11. Spontaneity and Subconscious Processing

12. Thinking and Evaluation

13. Value

14. Variety, Divergence and Experimentation

Out of these 14, some provide interesting relevance in the domain of stand-up comedy. “Domain Competence” is critical, as a comedian must understand the linguistic and cultural rules of the genre to subvert them effectively. “Social Interaction and Communication” is also vital, given that stand-up is inherently a dialogue with an audience where feedback (laughter) is immediate.

Due to the heavy emphasis on developing, workshopping and tightening a stand-up comedy routine discussed in Section 2.5 of the Background chapter, the focus of this thesis will be on “Progression and Development”. Jordanous and Keller define this component as the “movement, advancement, evolution and development during a process” (Jordanous and Keller 2016). They note that this progress is not necessarily linear and that the end goal may be only loosely specified.

This component provides the justification for shifting the research focus from “generation” to “refinement”. As supported by discussion in Section 2.5 of the Background chapter, a joke is rarely static and undergoes a rigorous feedback loop of Generate → Perform → Fail → Analyse → Rewrite. It evolves from a rough idea to a polished bit through repeated performance and refinement (the “tightening” of a set).

In prior work around computational humour, the systems often generate single jokes and stops, only exhibiting a fraction of the creative process. Evaluating a system on the “Progression and Development” component shifts the focus from a single static output (a one-off joke) to the system’s ability to refine, edit, and improve its material over time, simulating the professional comedian’s workflow.

The Engagement-Reflection Framework

While Jordanous and Keller focus on the evaluation of progress, Rafael Perez y Perez proposes a cognitive model for the generation of creative artifacts, specifically within storytelling and narrative. This Engagement-Reflection Framework posits that creativity is not a single continuous flow, but a cycle of two distinct states (Pérez y Pérez 1999).

- **Engagement:** The system generates material freely, driven by constraints and associations but without explicit critical judgment. This is where the model produces content based on learned patterns and immediate context.
- **Reflection:** The model pauses to evaluate the generated material against coherence requirements, novelty and interesting-ness. It guides the next stage of the “engagement” state.

This framework is particularly relevant to Large Language Models in comedy. A standard LLM generation can be seen as pure “Engagement” (a continuous prediction of the next result). However, “Reflection” is missing unless explicitly engineered through techniques like chain-of-thought or iterative refinement, where the model is prompted to critique its own joke before finalising it.

3.1.2 Perspectives on Computational Creativity Evaluation

Beyond specific metrics like novelty and value, the evaluation of Computational Creativity is often categorised into four distinct perspectives, known as the “Four Ps”: Person, Process, Product, and Press (Rhodes 1961).

- **Person:** Focuses on the internal traits of the creative agent (human or machine), such as “Skill”, “Appreciation”, and “Imagination”. This is famously formalised in Colton’s “Creative Tripod” framework, which argues that for a system to be perceived as creative, it must appear to possess these three

qualities (Colton 2008). Since LLMs are arguably devoid of personality, current research often resorts to ‘persona prompting’ to simulate the traits of a job or person.

- **Process:** Evaluates the creativity of the methods used to generate the artifact. This perspective asks whether the system explores a conceptual space in a transformative way, rather than just following a fixed algorithm (Goes et al. 2016).
- **Product:** The most common form of evaluation, focusing solely on the output (e.g., the joke itself) using metrics like those proposed by Ritchie (‘such as typicality and quality’) (Ritchie 2007).
- **Press:** Examines the interaction between the creative agent and its environment or audience. In the context of stand-up comedy, the “Press” is the audience, and their reaction (laughter) is the ultimate validator of the creativity.

The “Readymade” / Conceptual Mash-Ups

Tony Veale’s work argues that for computers to achieve genuine wit (a domain very linked to stand-up comedy), they don’t need to simulate deep human consciousness but rather exploit the rich, associative messiness of language. Veale introduces the concept of the “Linguistic Readymade”, found objects recontextualised as art (Veale 2013).

Veale suggests that creativity often acts as a retrieval and decision process rather than a pure construction process. He argues that the web acts as a “vast and growing treasury of readymade expressions” that systems can harvest. To bring this idea into the modern LLM era, the web refers to the LLM’s dataset which contains many of these “readymades”. By retrieving “found” linguistic objects, such as idioms, cliché’s or stereotypes, and recontextualising them, a

system can produce outputs that appear deeply creative without requiring deep world understanding (Veale 2012).

This leads to his theory of “Conceptual Mash-Ups”. While traditional Conceptual Blending requires complex mapping, Veale’s mash-ups are “computationally feasible” blends. They validate novel combinations not through logic, but through “Google milking”, a process of querying search engines to find linguistic proof that a concept is culturally resonant. If a system wants to know if “poets” and “poverty” are linked, it does not check a knowledge base;. Rather it queries the web for phrases like “poets are poor” or “starving poet” (Veale, Feyaerts and Forceville 2013). This validates the “cultural truth” of a joke setup without needing an ontology of truth.

This presents an interesting idea in stand-up comedy, as any written material needs to be culturally relevant and recognisable by an audience. Veale’s work provides an alternative framework through which to evaluate a stand-up comedy routine, regardless of authorship.

P-Creativity vs. H-Creativity

Finally, it is important to distinguish between P-Creativity (Psychological) and H-Creativity (Historical). Boden argues that a system displays P-Creativity if the idea is novel to the agent itself, even if it has been done before by others. H-Creativity, by contrast, requires the idea to be novel in the entire history of the domain (Boden 2004).

For an LLM generating stand-up comedy, achieving H-Creativity is exceptionally difficult due to the “hallucinated biography” problem (the model cannot truly invent a new human experience) (Mirowski et al. 2024). However, P-Creativity is a viable metric for success. If the system can identify a premise and, through an iterative process, refine it into a state that is novel (relative to its starting point), it has demonstrated a form of psychological creativity essential for the “Progression

and Development” component defined by Jordanous and Keller (Jordanous and Keller 2016).

3.1.3 Summary

Traditional definitions, such as Ventura’s requirement for intentionality or Ritchie’s criteria for strict novelty, are difficult to apply to large language models. Due to LLMs being probabilistic engines that are trained on vast samples of human knowledge, proving they have ‘intent’ or that an output is strictly ‘new’ (not present in the training data) is philosophically and computationally intractable.

However, by shifting the lens from Product (the final material) to Process (the refinement loop), a clearer path for evaluation emerges from this discussion. Veale’s “Readymades” provide the mechanism: acknowledging that creativity in this domain is often about the re-contextualisation of existing cultural artifacts. Jordanous & Keller’s “Progression and Development” provides the focus for creative reflection, moving away from static generation toward a dynamic system that can refine its own material. Boden’s “P-Creativity” shows the scope: evaluating the system not on whether the outputs are novel in the history of comedy (H-Creativity), but on whether it can successfully evolve its own internal concepts.

These ideas of creativity in computational systems supports this research that focuses on iterative refinement. Rather than treating the LLM as a ‘joke-machine’ (pulling the handle to get a joke), this thesis will treat the LLM as a ‘curator’ and ‘reflective practitioner’ evaluating its ability to engage in the feedback loop of drafting, failing, analysing, and rewriting.

3.2 LLMs and Prompting

To understand how Large Language Models (LLMs) can be applied to computational comedy, it is essential to first understand their underlying architecture.

Modern LLMs, such as OpenAI’s GPT series or Google’s Gemini, are built upon the Transformer architecture, first introduced by Vaswani et al. in 2017 (Vaswani et al. 2017). Unlike earlier Recurrent Neural Networks (RNNs) that processed text sequentially, Transformers process entire sequences of text simultaneously.

The core innovation of the “Transformer” architecture that underpins these models, is the “self-attention” mechanism. This allows the model to weigh the contextual importance of every word in a sentence relative to every other word, regardless of their physical distance in the text. In the context of comedy, this is what enables the model to connect a punchline back to a distant premise, facilitating the structural necessity of core comedic concept, the callback.

LLMs undergo a two-step training process. The first is “Pre-training,” where the model is exposed to vast, terabyte-scale corpus of internet text. During this phase, it learns grammar, facts, reasoning abilities, and structural patterns by constantly predicting hidden words in a sequence. The second phase is “Fine-tuning,” which often includes ‘Reinforcement Learning from Human Feedback (RLHF)’. In this stage, human evaluators rank the model’s responses, teaching it to align with human preferences for helpfulness, coherence, and safety (Ouyang et al. 2022). The technology functions as a highly sophisticated statistical engine, mapping high-dimensional vector relationships between concepts to generate text that mimics human reasoning and creativity.

3.2.1 Using LLMs in creative writing

To understand the capabilities and limitations of Large Language Models (LLMs) in the domain of stand-up comedy, it is important to define their fundamental operational mechanism: autoregressive next-token prediction. At their core, models such as GPT-4 are probability engines designed to predict the likelihood of a specific token (a word or part of a word) appearing next in a sequence, given the context of the preceding tokens (Brown et al. 2020).

Mathematically, the model seeks to maximise the conditional probability $P(w_t|w_1, \dots, w_{t-1})$, where w_t is the next token and w_1 to w_{t-1} represents the context window or the ‘prompt’ provided by the user.

This means that when an LLM writes a joke, it is not ‘thinking’ about a funny concept or planning a punchline in the way a human comedian does. Instead, it is calculating which word statistically tends to follow the previous words based on the billions of text parameters it processed during training (Brown et al. 2020).

This mechanism presents a unique paradox for humour generation. As established in Section 2.2.1 of the Background chapter, comedy relies on ‘incongruity’ and the “violation of expectation” (Script Opposition). However, an LLM is optimised to generate the most plausible continuation of a sentence. This is, effectively, the least surprising outcome. If a model strictly follows the path of highest probability, it will likely produce a statement that makes perfect semantic sense but lacks the necessary ‘twist’ to be funny (Toplyn 2023).

This challenge that researchers have needed to overcome of how the model aggregates the collective ‘average’ humour of its training data, resulting in low level open-mic comedy rather than unique artistic insight, has been addressed in a few different ways. Researchers and users manipulate parameters such as ‘temperature’, which introduces stochasticity (randomness) into the selection process. By raising the temperature, the model is permitted to choose less probable tokens, thereby increasing the chance of generating the novelty required many approaches to identifying Computational Creativity in systems (Ventura 2016).

However, this comes with the risk of ‘hallucination’ or incoherence, where the output becomes purely nonsensical rather than humorously absurd. Therefore, using LLMs for stand-up comedy becomes a delicate balancing act of prompt engineering: guiding the model away from the boring statistical average without pushing it into chaotic randomness (Li et al. 2023).

3.2.2 Prompt Engineering

While next-token prediction drives the atomic generation of text, the quality of complex creative outputs is significantly influenced by the structure of the interaction (prompt), through a discipline known as prompt engineering. Recent research indicates that for multifaceted tasks (which would include writing stand-up routines which require many steps, such as premise generation, structural formatting, and stylistic refinement) a single prompt is often insufficient (Sun et al. 2023).

To address this, researchers have developed ‘Prompt Chaining’, a technique that breaks a complex task into a sequence of discrete sub-tasks. In this workflow, the output of one prompt becomes the input for the next. For example, a “Generator” prompt might produce a raw list of ideas; a second “Critique” prompt then evaluates these ideas against specific criteria; and a third “Refiner” prompt rewrites the selected ideas into a final draft (Sun et al. 2023) as shown in Figure 2. This method mirrors the iterative ‘Draft $\rightarrow$ Analyse $\rightarrow$ Rewrite’ cycle of human comedians established in Section 2.5 of the Background chapter, allowing the model to ‘self-correct’ and guide itself before finalising the output.

By using this chain, it is effectively simulating the “Process” evaluation mentioned by (Goes et al. 2016). Instead of a ‘fixed algorithm’ that spits out a joke in one go, the system is forced into a step-by-step exploration of the conceptual space.

An alternative approach is “Stepwise Prompting,” which attempts to integrate these phases (drafting, critiquing, and refining) within a single, complex instruction, as show in Figure 3. However, studies suggest that Prompt Chaining typically yields more favourable outcomes for reasoning-heavy tasks. This is because chaining reduces the cognitive load on the model at any single step, preventing it from losing track of instructions or hallucinating constraints (Sun et al. 2023),(Li et al. 2023). However, it is reliant upon the internal memory of the model i.e. how much

- **The Generator (Drafting)**

Example prompt from ‘Generating’ step in Prompt-Chaining

Generate 5 short premises for a stand-up bit about the experience of being a ‘digital nomad.’ Focus on the contrast between the Instagram version and the reality of working from a laptop in a laundry mat.

- **The Critic (Analysing)**

Example prompt from ‘Critic’ step in Prompt-Chaining

Act as a cynical comedy club booker. Evaluate these 5 premises. Identify which one feels too much like a ‘hallucinated biography’ (a cliché the AI thinks is true) and which one has a specific, relatable observation. Rate them for Typicality and Quality.

- **The Refiner (Rewriting)**

Example prompt from ‘Refiner’ step in Prompt-Chaining

Using the feedback from the booker, rewrite the highest-rated premise into a 200-word monologue. Focus on sensory details (the smell of the laundry detergent, the glitchy Wi-Fi) to ground the ‘Person’ perspective and make the bit feel lived-in.

Figure 2: Example prompt, applying concepts of ‘Prompt Chaining’, specifically ‘Generator’, ‘Critic’ and ‘Refiner’

knowledge of previous responses it carries forwards into its next response.

‘Stepwise’ Prompting Example

Example prompt applying ‘Stepwise Prompting’

Act as a comedy writer and complete this task by following these exact steps in order. Please show your work for each step.

Step 1 (Drafting): Generate 5 short premises for a stand-up bit about the experience of being a ‘digital nomad.’ Focus on the contrast between the Instagram version and the reality of working from a laptop in a laundromat.

Step 2 (Analysing): Switch personas to a cynical comedy club booker. Evaluate your 5 premises from Step 1. Identify which one feels too much like a ‘hallucinated biography’ and which one has a specific, relatable observation. Rate them for Typicality and Quality.

Step 3 (Rewriting): Using your evaluation from Step 2, rewrite the highest-rated premise into a 200-word monologue. Focus on sensory details (the smell of the laundry detergent, the glitchy Wi-Fi) to ground the ‘Person’ perspective and make the bit feel lived-in.

Figure 3: Example prompt, applying concepts of ‘Stepwise Prompting’

Zero-shot vs. Few-shot

A critical variable in prompt engineering is the amount of context or examples provided to the model within the prompt itself. This is generally categorised into “zero-shot” and “few-shot” learning.

“Zero-shot” prompting involves presenting the model with a task without providing any examples of the desired output. In this mode, the model relies entirely on its pre-trained parametric knowledge (the data it ingested during its initial training phase) (Brown et al. 2020). While this demonstrates the model’s raw generalisation capabilities, it often leads to generic or ‘average’ outputs because the model defaults to the most statistically probable response found in its training data (Kharawala et al. 2024). An example of this type of prompting is shown in Figure 4.

In contrast, “few-shot” prompting involves providing the model with a set of high-quality examples (demonstrations) within the prompt before asking it to perform the task. For instance, providing three examples of lines of poetry before asking the model to generate a fourth. This technique allows the model to “recognise the pattern” of the specific task and adapt its output style without the need for updating its underlying weights (Brown et al. 2020). An example of this is type of prompting is shown in Figure 5.

- **‘Zero-Shot’ Prompting example**

Example prompt applying ‘Zero-Shot Prompting’

Write a short premise for a stand-up ‘bit’ about the experience of being a ‘digital nomad.’ Focus on the contrast between the Instagram version and the reality of working from a laptop in a laundromat.

Figure 4: Example prompt, applying concepts of ‘Zero-Shot’ Prompting

In the domain of stand-up comedy, ‘few-shot’ prompting is particularly valuable for style transfer, which is the ability of a model to mimic a specific comedian’s voice, rhythm, and persona. However, poor examples can lead to “negative transfer,” where the model replicates the flaws of the prompt rather than the intent (Brown

- **‘Few-Shot’ Prompting Example**

Example prompt applying ‘Few-Shot Prompting’

Task: Write a short premise for a stand-up comedy bit about the experience of being a ‘digital nomad,’ contrasting the Instagram version with reality.

Example 1:

Premise: ‘‘Everyone thinks I’m working from a pristine beach in Bali. The reality is that sand is destroying my MacBook’s fan while I try to find a public bathroom without a cover charge.’’

Example 2:

Premise: ‘‘My Instagram says #VanLife and shows me doing yoga at sunrise. My lower back says #ChiropractorNeeded because sleeping in a converted 2004 Ford Transit is just aesthetic homelessness.’’

Now, generate a new premise following this style:

Figure 5: Example prompt, applying concepts of ‘Few-Shot’ Prompting

et al. 2020). Even though providing examples acts as a “style guide” and could significantly reducing the “Cold Start” problem where the model struggles to understand the specific persona or tone required (Kharawala et al. 2024) producing original content needs to be the priority. Relying heavily on examples inhibits the “originality” component of the creativity framework proposed by Jordanous and Keller (Jordanous and Keller 2016), or similarly the “novelty” aspect of Ventura’s criteria for Creativity (Ventura 2016).

This distinction is critical for stand-up comedy because the genre’s success relies on a delicate balance between typicality and novelty. While ‘few-shot’ prompting effectively solves the “Cold Start” problem by giving the AI a recognisable character, persona or voice, it simultaneously risks turning the model into a high-tech impressionist rather than an original creator. In comedy, a joke that perfectly mimics the style of a famous comedian but offers no new insight is often dismissed as a “pastiche” or a “hack” routine. To achieve the previously mentioned H-Creativity, the system must eventually break away from the examples it’s given.

3.2.3 Methods of Prompt Engineering

Beyond the structural sequencing of ‘Prompt Chaining’, there are different methods that have developed regarding “How to write a prompt”. It is suggested that these methods help to improve the output of an LLM model, by helping it to navigate its training data. Prompt Engineering started as a simple trial-and-error process and has developed into formalised strategies designed to maximise the capabilities of the transformer architecture.

Framework-based Prompting

To avoid the ambiguity that can occur in natural language, researchers have adapted traditional frameworks for generative AI. Two prominent examples of this are the SMART and CLEAR strategies.

The SMART framework (adapted from management theory) suggests that a prompt must be “Specific, Measurable, Action-oriented, Relevant, and Time-bound” (or Task-bound). For example, instead of asking for “something funny,” a SMART-aligned prompt specifies the exact persona (Specific), the number of jokes required (Measurable), the specific comedic device to use, such as a “call-back” (Action-oriented), the target demographic (Relevant), and the word count limit (Task-bound) (Saunders and Gupta 2024).

Similarly, the CLEAR framework emphasizes the linguistic quality of the instruction: “Concise, Logical, Explicit, Adaptive, and Reflective”. This method is particularly effective for navigating the ‘context window’ of a model. By being Explicit about what to avoid (e.g., “do not use puns”) and Reflective by asking the model to explain its choice of punchline, the prompt engineer can achieve a higher degree of technical precision (Zhao, Patel and Rossi 2023).

These frameworks present complementary approaches to prompt optimisation, differing in their focus on output and linguistics. The SMART framework focuses on constraints and objectives, ensuring that the outputs from a model meets the criteria. In contrast, the CLEAR framework focuses on the internal logic and clarity of the instruction, removing ambiguity and helping the model to better ‘understand’ the task.

While SMART ensures the model produces a routine of a specific length (Measurable), CLEAR ensures the model avoids low-quality tropes and can justify its narrative choices (Explicit and Reflective). Ultimately, this research will bring together aspects from both the SMART and CLEAR frameworks to develop highly specific, crafted prompts tailored to the nuanced requirements of stand-up comedy generation and the reflective process that is being built.

Framework	Criteria	Example
SMART	• S: 5 taglines for "The Gilded Crust" • M: 3–6 words per tagline • A: Focus on sensory texture • R: Luxury demographic • T: Numbered list + justification	Act as a branding consultant. Generate 5 taglines for 'The Gilded Crust,' a luxury cheesecake bakery. Each must be between 3 and 6 words and emphasize the silky texture of the product. The tone should be sophisticated for a high-end market. Provide the results in a numbered list with a brief positioning statement for each.
CLEAR	• C: Single impactful tagline • L: Link heritage with modern luxury • E: No "creamy", "sweet", or puns • A: Minimalist, elegant tone • R: Reflective explanation	Generate a single tagline for a premium cheesecake restaurant. The phrase must logically connect artisanal heritage with modern luxury. You are strictly forbidden from using the words 'delicious,' 'creamy,' 'sweet,' or any cheesecake-related puns. Use a minimalist, sophisticated tone. Finally, justify your linguistic choices and explain how they avoid the 'generic' pitfalls of food advertising.

Table 1: A comparison of SMART and CLEAR prompt frameworks

Persona and Role-based Prompting

A core method in creative prompt engineering is the establishment of a ‘persona’. Rather than treating the model as a generic information retrieval system, persona prompting assigns the AI a specific identity, professional background, and set of behavioral constraints. This technique exploits the model’s ability to bring together relevant traits found in its training data. For instance, assigning the role of a “1980s observational comic” versus a “modern surrealist” drastically shifts the model’s language choices, pacing, and thematic focus (Miller 2024). This will be discussed further in the Methodology chapter of this thesis.

3.2.4 Summary

This research highlights some distinct challenges that may be faced in comedy generation using LLM technology. While the model is optimised for plausibility and coherence, humour relies on incongruity and the violation of expectation. A model that strictly follows the path of highest probability will produce text that is semantically sound but ‘unfunny’ and predictable. While manipulating parameters like ‘temperature’ can introduce the stochasticity required for novelty, it risks pushing the output into incoherence or hallucination. (Ventura 2016).

The application of framework-based and persona-driven prompting serves as a suitable intervention to this problem of predictability. By utilising aspects of the SMART and CLEAR frameworks, the prompt engineer moves beyond trial-and-error to enforce the technical precision and linguistic constraints necessary to avoid generic tropes. Furthermore, by grounding the model in a specific persona, a prompt engineer can leverage the model’s latent training data to adopt the unique pacing and vocabulary of a professional performer. These methods provide the qualitative ‘DNA’ for the prompt, ensuring the model navigates its state space toward intentional results rather than mere statistical probability.

To further mitigate these limitations, the literature suggests that the structure

of the prompt is as critical as the model itself. For complex creative tasks like stand-up, ‘Prompt Chaining’ (where the output of one step becomes the input for the next) is superior to single-turn ‘Stepwise Prompting’ (Sun et al. 2023). By breaking the task into discrete phases of “Generation”, “Critique”, and “Refinement”, researchers can reduce the model’s cognitive load and simulate the iterative ‘Draft → Analyse → Rewrite’ cycle employed by human comedians.

3.3 AI in the domain of Humour

There is a distinct gap in the literature when tackling humour using artificial intelligence. There has been quite a lot of exploration and research around one-liner short-form comedy, improv and comedy sketches, whereas there has been less research done to tackle the creation of extended (longer form) stand-up comedy routines using this technology.

3.3.1 Computational approaches to generating humour

Rule Based Approaches

The history of computational humour generation is broadly categorised into two distinct eras: the symbolic (rule-based) approach and the neural (stochastic) approach. Early research, often termed “Foundational Automators,” relied heavily on symbolic processing, where the system was explicitly programmed with linguistic rules and fixed templates (Amin and Burghardt 2020). Weizenbaum’s “ELIZA” was not humour-focused but laid the foundation for these systems. The true shift toward generative humour began with the formalisation of jokes as some kind of “solvable linguistic puzzle”. These systems, such as “JAPE” (Joke Analysis and Production Engine) developed by Binsted and Ritchie (Binsted and Ritchie 1994), utilised rigid formulas and lexical databases like ‘WordNet’ to manufacture jokes. The advantage of this approach was structural guarantee; because the system

filled in a pre-defined ‘slot-and-filler’ template (e.g., “What do you call a X that is Y?”), the output always adhered to a recognisable joke format (Winters, Nys and De Schreye 2018). However, these systems were severely limited by their reliance on hard-coded logic, resulting in humour that was mechanically repetitive and lacking in semantic depth.

There were several distinct systems that paved the way for modern generative humour. A key early example is “HAHAcronym”, a system designed to generate humorous re-interpretations of existing acronyms. By replacing words in an acronym with antonyms or semantically incongruous alternatives while preserving the original letters, the system demonstrated that humour could be generated via lexical substitution constraints without a ‘sense of humour’ module (Stock and Strapparava 2003).

Later systems moved beyond simple substitution. “Metaphor-Eyes” utilises Veale’s “Google milking” technique to generate creative metaphors by projecting stereotypical properties from a source domain to a target domain. For instance, it can metaphorically describe “Google” as a “cult” by verifying that they share linguistic associations (e.g., “gurus,” “followers”) in web corpus (Veale 2014).

Tony Veale’s later work highlights that machines are not inherently unfunny, but rather “programmed that way” due to a lack of world knowledge (Veale 2021). Veale argues that for a system to possess a ‘sense of humour’, it must treat joke generation as a knowledge-based process, blending “wit and wisdom” rather than just shuffling linguistic patterns. His work on metaphor and irony demonstrated that computers could generate “witty” comparisons by mapping properties between disparate concepts, a precursor to the high-dimensional vector mapping seen in modern “Transformer” models.

Early Neural Models and Transformers

The paradigm shifted significantly with the advent of Neural and Statistical models, specifically “Recurrent Neural Networks” (RNNs), “Long Short-Term Memory networks” (LSTMs), and eventually “Transformers”. The release of “Sunspring”(Sharp and Goodwin 2016), an AI-written screenplay generated by an LSTM network demonstrated that neural models could produce humour, although largely unintentional at this point.

Unlike their symbolic predecessors, these models do not rely on explicit linguistic rules but instead ‘learn’ humour by training on vast corpora of existing jokes (Amin and Burghardt 2020). This allows for “Open Domain” generation, meaning the AI is not restricted to specific topics or templates. However, while neural models excel at linguistic fluency, they often struggle with the ‘logical mechanism’ of a joke. As noted by Winters, deep learning models often produce text that looks like a joke (surface structure) but lacks the meaningful incongruity (deep structure) required to be funny, often resulting in nonsensical or hallucinated punchlines (Winters, Nys and De Schreye 2018).

Approaches to Text generation

In the lead-up to comedy generation, we saw the use of GPT-2 for creative text generation prior to RLHF. Researchers fine-tuned GPT-2 on specific corpora (e.g., Romantic-era poetry or monologues) to induce a ‘stylistic transfer’ (Sawicki et al. 2022). However, these models often struggled with “long-range coherence,” frequently contradicting their own setups, which is a fatal flaw for being able to generate stand-up routines that rely on callbacks and consistent personas (Brown et al. 2020).

Due to the structural rigidity of early symbolic systems, the majority of early success in computational humour has been in the domain of puns and riddles. Puns are computationally tractable because they rely on “phonological

ambiguity” (a quantifiable feature where words sound the same but have different meanings) (Toplyn 2021). Systems like “HAHAcronym” and the “STANDUP” project (designed for children with cerebral palsy) successfully generated thousands of punning riddles by simply traversing lexical graphs to find phonetic matches (Amin and Burghardt 2020). In these short-form jokes, the setup is often just a definition, and the punchline is the homophone, requiring very little world knowledge or narrative reasoning. Although successful, this type of humour is quite distinct from the stand-up comedy goals that this research is exploring.

Transitioning to narrative-driven jokes and stand-up routines presents a far greater challenge, described by Toplyn as “AI-Complete” (Toplyn 2022). Narrative jokes require the system to maintain a “coherent story-line” while simultaneously planning a subversion of that story. While “Transformer” models (like GPT-3.5 and GPT-4) have shown the ability to generate anecdotal text, some fail to execute the specific setup-punchline arc because the function of the model (predicting the most probable next word) is at odds with the objective of a joke (providing a surprising, low-probability conclusion) (Peyrard and West 2021). Consequently, current research often employs hybrid systems, combining the narrative fluency of Large Language Models with the structural constraints of symbolic algorithms to attempt to bridge the gap between simple wordplay and complex storytelling. This is the basis for Toplyn’s most recent “Witscript 3” system. (Toplyn 2023). This bridges some of the gap between the downsides of symbolic systems but still runs into predictability and lack-of-funny.

3.3.2 Computational Performance

There have been a number of examples of performers and academics taking AI onto the stage, in improvised and/or comedy performances. In a lot of these examples the AI acts as an ‘unpredictable’ co-performer.

Improbabilities team Kory Mathewson, Riley Dawson and Piotr Mirowski have

taken to the stage using Artificial Intelligence and Robots. Mathewson, named one of Alberta's "Top 40 Under 40" for his work in Human-AI interaction (Barlott 2018), explored the concept of "Artificial Improvisation". The "Blueberry" Robots and artificial intelligence was trained on the dialogue adapted from hundreds of Hollywood movies. This robot, running software called "ALEX", tested their creation in front of a live audience at Edmonton Fringe. This show was based around testing if "Blueberry" can improvise for an audience and interact with different audience members. According to Mathewson et al., the show went "incredibly well" and the "critical review was very positive" (Mathewson and Mirowski 2017). Oliver Moody notes the challenge of "contextual reverence" where the comedy often arose not from the AI's cleverness but from its failure to adhere to social norms (dubbed "Artificial Irreverence") (Moody 2021). The show, including the "Blueberry" robot, demonstrated that audiences could engage with an AI that made mistakes, provided those mistakes were framed within the "Yes, And" contract of improvisation.

In Colleen Lavin's Edinburgh Fringe Show "Do The Robots Think I'm Funny?" (Girdwood 2023) she used her "Murderbot" system to judge her jokes and determine whether she is funny, based on the laughter she obtained from the audience. This system then heckled her when she didn't get laughs. The "Murderbot" system tracks the audience response, specifically laughter, and if it detects a too-long pause, it will heckle Lavin. The audience has the option to "boo" the Murderbot if they believe that the heckle was unwarranted.

Philipp Wicke has explored robotic stand-up, alongside Veale, investigating how embodied agents can deliver humour. The work highlights that a system must do more than generate text; it must understand the gesture and timing that accompany the delivery. They argue that an AI comedian must effectively 'perform' the script, using physical cues to signal the punchline, much like a human comic (Wicke and Veale 2021).

In the text “Your Wit Is My Command”, Veale synthesises these ideas, proposing that we are moving toward systems that are not just “generators” but “curators” of wit-machines that understand the social dynamics of a joke, including sarcasm and irony, rather than just the linguistic structure (Veale 2021).

More recently, investigations have been done within the physical domain. While the earlier work focused on the robot’s physical performance, this recent research addresses whether AI models can actually understand the “Image Schemas” (the mental maps of space and movement) that underpin human communication.

The focus has shifted from simple text generation to spatial reasoning, evaluating whether LLMs can ‘see’ a performance in their mind’s eye. Does a model understand that a “shrugging” gesture conveys the same semantic uncertainty as the words “I don’t know”. (Wicke and Wachowiak 2024).

3.3.3 AI & Comedians

As mentioned, late-night comedy show writer Joe Toplyn has contributed “Witscript”, a system for creating “improvised jokes in conversation”, “Witscript 2”, for creating jokes using wordplay and “Witscript 3”, a (claimed) more advanced system. These systems combine his theory of joke writing, aptly named the “Surprise Theory of Laughter”, his “Basic Joke-Writing Algorithm” and GPT-3. The results were “encouraging because it shows that “Witscript 2” can create jokes that are as successful as, but more sophisticated than, mere word-play jokes” (Toplyn 2022).

Although not using LLMs, in 2021, Netflix published their first bot-generated stand-up routine with help from Keaton Patti (Patti 2021). As cited by the viewers, “50% of the jokes are actually pretty funny”. It appears that the general consensus was that most viewers could find a joke(s) that they found funny. However, the validity of it has later been brought into question, and how much of the content was computationally generated.

After the public release of ChatGPT, from OpenAI, public figures began to use

these models in comedy projects. A YouTube musical comedian, Daniel Thrasher, used ChatGPT to generate comedy sketches through a series of self-written prompts (Thrasher 2023). The audience reaction was mixed. Many praised what the video was able to achieve, stating impress at how the system could create a sketch with a “coherent story-line”, but others thought that Daniel’s usual content was far “funnier than the AI”, despite the expressive performance and engaging editing. Some viewers identified that the content being AI-generated made it funnier.

Comedian and late-night talk show host, Pete Holmes was asked to analyse ChatGPT-generated stand-up comedy in an article by Vulture. He described the routines generated as “it’s writing what I would consider very standard open-mic-level comedy” (Pandya 2023). This provides a useful time-stamp from a domain-expert as to the abilities of LLM models within this domain. Academic Oliver Double also argues that AI cannot replicate stand-up comedy because it lacks a “persona, risk, and the ability to read a room”. This limitation may have steered research toward “Creativity Support Tools” rather than autonomous replacements. The DeepMind study “A robot walks into a bar” (Mirowski et al. 2024) and a review of ChatGPT and Google Gemini comedy by Techradar (Wilson 2024) has documented comedians using AI, not as a writer, but as a brainstorming partner.

3.3.4 Summary

This literature covered in this section establishes a clear progression in computational humour, moving from rigid, template-based symbolic systems (like “JAPE” and “HAHAcronym”) to the fluid, open-domain capabilities of Neural Models (like GPT-4) which is where this research is positioned. However this review identifies a critical gap between short-form linguistic puzzles and long-form stand-up routines.

Current research has various progress in humour based on phonological ambiguity. Systems can easily traverse lexical graphs to find homophones, as seen in

the success of “Witscript 2” and “STANDUP”. However, there is a lack of research addressing the generation of cohesive stand-up sets.

Furthermore, this review highlights that current successful implementations are predominantly collaborative rather than autonomous. In projects like Improbotics the AI functions as an unpredictable co-performer to react to a human. The challenge that this research may face is reducing the reliance on human-collaborator to try to explore what the capabilities of large language models are. The question is: How can an AI not only generate material but ‘audit’ its own output, critique it against comedic principles and refine it to the polished routine? In order to conclude the success of any system at addressing this question, methods for evaluating comedy are needed.

3.4 Evaluating Humour

In most cases of demonstrating comedy by LLMs or performing ‘known’ comedy material to a public audience, it has become hard to identify the cause of the humour. Is the humour caused by well-written humorous material, the comedian’s performance and delivery, or just the fact that the comedy is produced by a large language model? Therefore detailed methods are required to evaluate humour produced by both humans and computational systems.

3.4.1 Audience metrics

The most empirical measure of comedic success is the physiological response of the audience: particularly laughter. In the domain of stand-up comedy, laughter functions as an immediate, binary signal of validation; a joke either “lands” or it “dies” (Smith 2018). However, academic analysis suggests that laughter is not a rigid metric but a complex spectrum. Researchers distinguish between “polite laughter” (social bonding), “agreement laughter” (validating a premise),

and “visceral laughter” (the involuntary response to incongruity) (Schwarz 2010).

In recent quantitative studies, such as the work by Cela et al., audience response is measured not just by presence but by intensity and duration. By correlating the “Laughs Per Minute” (LPM) or the decibel level of applause with specific linguistic markers (using tools like LIWC-22), researchers can objectively map the “peaks” of a routine (Cela, Veit and Wood 2024). Yet, this metric remains fundamentally subjective to the specific audience present. A routine that generates raucous applause in a progressive urban comedy club may face silence in a conservative rural venue, highlighting the difficulty of using audience reaction as a universal “ground truth” for training generalisable AI models (Wawrzyniuk 2021).

As seen in Improbotics’ work, the evaluation extends to the ‘cohesiveness’ of the scene. Does the AI maintain the “Yes, And” contract, or does it break the narrative flow? (Mathewson and Mirowski 2017). Mathewson and Mirowski argue that standard quantitative metrics for dialogue systems are insufficient for judging improvisational theatre because they fail to capture the “audience-perceived performance level” or the artistic success of the interaction. Instead, evaluation often relies on whether the system can adhere to the fundamental improvisational rule of “Yes, And”, the ability to accept a partner’s offer and expand upon it.

Furthermore, Mathewson and Mirowski note that while the system can generate immediate dialogue, it is not trained to tell a “cohesive story with a narrative arc” or utilise memory for callbacks, which are critical for long-term coherence. Consequently, evaluating AI in an improvisational domain requires looking beyond the text to the “stage presence” and the system’s ability to act as a collaborative teammate rather than just a text generator. However, as this work measures the abilities of LLMs to construct routines and not the performance of human comedians performing AI-generated material, or AI-driven voice generators, different metric systems are needed for evaluating humour without performing it to an audience.

3.4.2 Non-audience metrics

In the absence of a live audience, researchers have turned to computational linguistics to quantify the structural properties of humour. These metrics attempt to measure the incongruity and resolution stages of a joke using statistical and vector-based measurements.

Perplexity and the ‘Surprise’ Curve

The most widely utilised metric in generative modelling is “perplexity”, which measures the uncertainty of a probability model in predicting the next token. In standard NLP tasks (like translation), lower perplexity is better. However, in humour generation, low perplexity equates to “predictability,” which is fatal for a joke (Vu et al. 2024).

Research by Peyrard et al. on Laughing Heads established that successful jokes exhibit a specific “Perplexity Signature” (Peyrard and West 2021). A standard joke structure consists of a setup with low perplexity (establishing a familiar script) followed immediately by a punchline with a local spike in perplexity (the incongruity). However, global perplexity measures are insufficient. If the perplexity is too high, the text is sometimes nonsensical from some models; if too low, it can be boring and predictable. Peyrard’s framework analyses “aligned pairs” of serious and humorous sentences, finding that Transformer models can detect humour by identifying specific tokens that disrupt the predictive flow of the sentence .

Semantic Distance and Vector Embeddings

To measure “Script Opposition” (as defined in the GTVH discussed in section 2.2.1 of the Background chapter), researchers utilise Semantic Distance. By mapping words into high-dimensional vector spaces (using embeddings like “Word2Vec”, “GloVe”, or “BERT”), systems can calculate the cosine similarity between the concepts in the setup and the concepts in the punchline (Kalloniatis and Adamidis

2024).

A framework proposed by Yang et al. (2015) suggests that humour resides in a “Goldilocks Zone” of semantic distance. If the distance is too short (< 0.2), the punchline is a logical tautology (no incongruity). If the distance is too large (> 0.8), the punchline has no resolution. The ideal joke connects two distinct semantic clusters that are distant enough to be surprising but close enough to be resolved via a ‘connector’ or logical mechanism (Kalloniatis and Adamidis 2024).

Attention Weights as Humour Detectors

Another more sophisticated framework involves analysing the “self-attention” mechanisms of the “Transformer” models. The Laughing Heads study demonstrated that specific ‘Attention Heads’ in models like “BERT” implicitly learn to identify the ‘disjunct’ (the keyword that triggers the ambiguity) (Peyrard and West 2021). By visualising the attention weights, researchers found that in humorous sentences, the model’s attention focuses intensely on the connection between the setup and the punchline, whereas in serious sentences, attention is distributed more evenly across grammatical structures. This suggests that ‘attention entropy’ can serve as a proxy metric for the presence of a Logical Mechanism.

Ambiguity and Phonological Features

For wordplay-based humour, metrics typically focus on “lexical ambiguity.” (Kao, Levy and Goodman 2016). Kao et al. argue that for a wordplay-based joke (like a pun) to work, the ‘process’ must create a state where two distinct meanings are equally plausible to the listener until the punchline ‘anchors’ them. The two measures they look at were ‘ambiguity’ and ‘distinctiveness’.

This analysis of “lexical ambiguity” involves counting the number of polysemous words (words with multiple meanings) using databases like “WordNet”. A framework reviewed by Amin & Burghardt utilises “phonetic similarity” scores to

evaluate puns, calculating the edit distance (Levenshtein distance) between the target word and the spoken word to ensure the pun is recognisable (Amin and Burghardt 2020).

While statistical metrics measure probability, heuristic frameworks measure structural competence. These are often rule-based filters designed to reject outputs that fail to meet the formal definition of a joke.

The ‘Witscript Internal Scorer’

Joe Toplyn’s “Witscript” system employs a rigid heuristic framework based on professional comedy writing theory. Before a joke is displayed to a user, it passes through an internal ‘filter’ that evaluates it against specific constraints:

- **Keyword Linking:** Does the punchline contain a keyword related to the “handle” (the focus) of the topic?
- **Angle Check:** Does the punchline take a different “angle” than the setup? The “angle” is a specific sub-topic or point of view derived from a broader subject
- **Length Constraint:** Is the punchline concise? (Longer punchlines dilute the surprise).
- **Question Prohibition:** Toplyn suggests that a punchline should rarely be a question, as this shifts the cognitive load back to the audience.

If a generated candidate fails these heuristics, it is discarded, regardless of its perplexity score (Toplyn 2023). This ensures that the system mimics the “tightening” process of a human comedian.

Toplyn’s framework provides insight because it moves beyond the probabilistic next-word nature of LLMs to enforce discipline of pre-defined comedic ideas. By prioritising structural constraints like “angle checks” and “length constraints”

over a model’s perplexity score, it demonstrates how a filter could mitigate the invented biography problem, where LLMs create a fabricated story, through rigid, human-derived heuristics.

Schema Learning and Template Matching

Winters et al. proposed a framework based on “Schema Learning.” Instead of judging a joke on its semantic content (which is subjective), this approach evaluates how well the generated text fits a known humorous structure, such as the “I like my X like I like my Y...” analogy template. The system is evaluated on its ability to fill the variable slots (X, Y, Z) with concepts that share specific relational properties (e.g., “dark” and “bitter” for coffee and ex-lovers). This framework shifts the evaluation from “Is it funny?” to “Is it a valid instance of this joke class?”, providing a quantifiable accuracy metric for structural generation (Winters, Nys and De Schreye 2018).

The work of Winters et al. provides a quantifiable metric by focusing on the structural integrity of a joke rather than its subjective funniness. This framework is highly useful as it offers a way to measure whether an LLM is correctly filling variable slots based on relational properties, rather than just outputting a statistically likely sentence. However, when applied to extended stand-up comedy routines, with varying structure, this template-based approach will be intractable.

‘JokeEval’ and Quality

A cutting-edge framework presented at ICML 2025, “JokeEval”, proposed a shift from evaluating the joke itself to evaluating the explanation of the joke (Jain 2025). This framework defines that if an LLM can successfully explain why a text is funny (identifying the incongruity and resolution), the text is more likely to be a valid joke. The framework uses a “crowd prompt” method where the AI assumes multiple personas to critique the joke’s logic. If the system cannot generate a

coherent explanation for the humour, the text is flagged as a failure/hallucination. This mimics the ‘Analyse’ phase of the human feedback loop, where a comedian must understand the mechanics of their own failure .

The “JokeEval” framework was a critical advancement because it uses meta-cognition to bridge the gap between a model’s output and its “intent.” By requiring the AI to successfully explain the incongruity and resolution, this method simulates the ‘Analyse’ phase of a human comedian, ensuring the humour is a result of logical design rather than a stochastic accident or a nonsensical hallucination. This is an idea that can be directly applied to any LLM-generation to better understand the ‘intention’ behind what the model has generated.

LLM-as-a-judge

This builds directly upon the idea of using LLMs to judge material. Preliminary work on using LLMs to assess the humour of jokes has been published, such as “Is GPT-4 Good Enough to Evaluate Jokes?” which was a poster presentation at the ICCV ‘23 conference. This paper explores using GPT-4 as a humour review model, by comparing it to human evaluators. The results show “a weak but encouraging positive correlation with human judges in the ranking of jokes compared to other simpler system descriptions” showing that LLMs have a type of ‘understanding’ of humour when given suitable data and direction (Góes et al. 2023).

Further validating the concept of using models for automated evaluation, Zheng et al. formalised the “LLM-as-a-judge” framework, demonstrating that strong LLMs (like GPT-4) can match or exceed human agreement levels when evaluating open-ended text generation (Zheng et al. 2023). Their research highlighted that while evaluating creative or subjective text traditionally required costly human annotation, appropriately prompted LLMs could consistently assess complex nuances such as tone, structural coherence, and alignment with instructions.

When applied to the domain of comedy, this suggests that an LLM does not

necessarily need a “sense of humour” to evaluate a joke. Instead, by leveraging its extensive memory and dataset of comedic theory and linguistic structures, it can function as a highly capable proxy judge. It can audit generated text against heuristic frameworks (such as identifying the presence of a setup, a logical mechanism, and a subversion of expectations) to score or critique the material without requiring a live audience.

3.4.3 Summary

These previous explorations have justified the decision of comedy writing as the genre of choice for testing. While current LLMs can mimic the linguistic structure of a joke or adhere to statistical probabilities that suggest humour, they lack a unified internal feedback loop.

There are different perspectives on how to evaluate humour. Heuristic systems like “Witscript” provide excellent structural guidelines but lack the semantic flexibility to generate novel, creative incongruity. Whereas statistical metrics like Semantic Distance measure the ‘surprise’ but fails to account for the ‘make sense’ aspect of the resolution. This can result in high surprise but also high surrealism rather than comedy. The use of ‘audience metrics’ provide a more truthful perspective, but are practically inaccessible during a prompt chaining workflow, making them unsuitable.

This identifies a clear area for research in the field: the transition from static generation to dynamic refinement. While Góes et al. (2023) demonstrated that LLMs possess a latent capacity to function as a “Judge”, and (Jain 2025) suggests they can generate ‘explanations’ for humour, there is limited research on closing this loop to automate improvement. The challenge of this thesis will be creating a workflow that can generate comedy, but also evaluate the material without audience metrics, to feed into an iterative refinement and improvement process.

3.5 Ethics in AI Comedy

The integration of Artificial Intelligence into the domain of stand-up comedy presents complex ethical questions that deserve acknowledge and discussion. These include discussions around the human-centric approach to stand-up comedy, the ‘nature of the self’ and the disembodied nature of Large Language Models (LLMs).

The fundamental philosophical friction at the heart of AI-generated stand-up comedy is the question of ‘lived experience’. As established in Section 2.1.3 of the Background chapter, contemporary stand-up is defined as an “autobiographical mode of performance” where the humour is linked to the performer’s identity and their specific location within the social hierarchy (Weaver and Lockyer 2024). Comedians do not just tell jokes. Rather, they “speak on their own behalf,” transforming their private tragedy or observation into public entertainment through a “performance of self” (Smith 2018).

Large Language Models, conversely, possess no ‘self’. They are ‘disembodied’ systems that generate text based on probabilistic correlations rather than phenomenological engagement with the world (Baeyaert 2025). When an AI generates a joke beginning with “I walked into a bar...” or “My wife is crazy...”, it is engaging in a form of “hallucinated biography.” While the audience in a human performance engages in a “suspension of disbelief,” accepting the comedian’s premise as a valid subjective truth, the audience for an AI knows that the premise is empirically false. This rupture in the ‘bonafide’ mode of communication threatens the ‘reliability’ that is often cited as the primary engine of successful observational comedy (Schwarz 2010).

The lack of lived experience becomes ethically critical when discussing the target of humour. Stand-up comedy functions as a ‘social corrective’ that often relies on ‘punching up’ (attacking the powerful) or ‘punching down’ (attacking the marginalised). Human comedians navigate these complex intersectional power dynamics using their own identity as a license. For example, a comedian from a

marginalised group can use self-deprecating humour to subvert stereotypes about their own community (Weaver and Lockyer 2024).

However, LLMs are trained on vast datasets of internet text which inherently contain historical biases, stereotypes, and prejudicial language (Brown et al. 2020). Because the model operates on statistical likelihood, it risks reproducing these stereotypes without the ‘ironic distance’ or ‘intent’ that a human comedian might employ to satirise them. The model does not understand the social weight of a slur or a stereotype; it only understands its semantic proximity to other words. Consequently, early experiments with AI humour have frequently resulted in outputs that are accidentally racist, sexist, or offensive, not because the AI is malicious, but because it mirrors the bias present in its training corpus (Jaokar, Arenal and Feijóo 2025).

To combat this, developers impose heavy “Reinforcement Learning from Human Feedback” (RLHF) to align models with safety guidelines, effectively censoring the system from generating ‘unsafe’ content. This creates a paradox for humour generation, which relies on the Benign Violation Theory discussed in Section 2.2.1. If a model is over-moderated, it removes the ‘violation’, resulting in comedy that is ‘safe, corporate, and unfunny’ often described by comedians as sounding like “a HR department wrote a joke” (Pandya 2023).

Comedian Gena B Jones made observations about a ChatGPT-generated joke in one of her sets, where she compared her punchline to the one given by the model. She identified that ChatGPT was censored by some kind of internal moderation and unable to make a joke about their topics. This identifies a clear challenge in stand-up comedy generation using AI as the model may filter out responses it deems ‘inappropriate’. Although not a limitation, a lot of stand-up comedy involves family friendly topics, it does vastly reduce the outputs that contain certain crude subject matters and the use of swear words and rudeness for emphasis.

The immense generative capabilities of Large Language Models have reignited

legal and ethical debates surrounding plagiarism and derivative content. Unlike traditional ‘copy-paste’ plagiarism, AI models engage in a high-level mimicry learning statistical probabilities of word associations from vast corpora to generate new text that feels familiar yet distinct (Khatiwada et al. 2024). This creates a ‘derivative work’ paradox. While the output may be syntactically novel, it often retains the semantic fingerprint of the training data, effectively washing copyright through a neural network (Jaokar, Arenal and Feijóo 2025).

In the context of stand-up comedy, where an individual joke is defined as much by its unique rhythm and perspective as its words, this raises the question of whether an AI mimicking a comedian’s specific style constitutes theft. Current intellectual property frameworks struggle to classify this ‘style transfer’ as infringement because copyright typically protects the expression of an idea, not the style itself (Khatiwada et al. 2024). However, Baeyaert (2025) argues that we must move toward a ‘provenance-aware’ ethics, where originality is not just about the final output but about transparently crediting the human artifacts that made the generation possible (Baeyaert 2025). Without this, AI comedy risks becoming a ‘stochastic parrot’ of existing copyrighted material, stripping the economic and moral value from the original human creators (Jaokar, Arenal and Feijóo 2025).

The use of AI raises significant ethical questions regarding ‘cultural ownership’. Styles of comedy are often deeply rooted in specific cultural traditions (e.g., the African American “dozens” or Jewish vaudeville). When an LLM is prompted to “write a joke in the style of [Comedian X],” it effectively performs a style transfer that strips the aesthetic from its cultural context (Khatiwada et al. 2024). This commodification of style, without the accompanying lived experience of the culture that produced it, risks becoming a form of appropriation, where the machine mimics the linguistic cadence of a community to which it does not belong.

The fear of displacement is palpable in the creative industries. In the documentary “Will a Robot Steal My Job?”, Anne-Marie Tomchak investigates the rise

of ‘artificially intelligent comedians’, noting that while machines can replicate the structure of a joke, they currently lack the nuance to ‘read the room’ (Tomchak 2019). However, the threat is not just replacement, but devaluation. As generative art floods the market, there is a risk of unique voice of the human comedian being drowned out by an infinite supply of statistically average content (Khatiwada et al. 2024). This leads to an ethical consideration, whereby as creatives using AI, it should be positioned as a collaborative tool (like a writing partner) rather than an autonomous replacement (Jaokar, Arenal and Feijóo 2025).

3.5.1 Ethical Implications of the Research

This research deliberately positions itself within these ethical boundaries by shifting the focus from autonomous generation to iterative refinement. By treating the LLM as a ‘curator’ and structural editor rather than an autonomous comedian, this thesis avoids the ethical pitfall of fabricating the lived experience explored by human comedians. The system is not being tested on its ability to invent a false, relatable persona, but rather on its mechanical ability to audit, critique, and improve the structural incongruity of a comedic premise.

Furthermore, by concentrating on the perspective of refining comedy through the “Progression and Development” component outlined by Jordanous and Keller (Jordanous and Keller 2016), this research champions the ‘collaborative tool’ model over the ‘replacement’ model. Instead of relying on ethically dubious style transfer prompts that appropriate the unique voices of human comedians, the methodology explores the fundamental, underlying mechanics of joke writing. This ensures that the AI is evaluated on its capacity to facilitate the creative process, acting as an analytical sounding board for the ‘Draft → Analyse → Rewrite’ cycle, rather than attempting to usurp the deeply human, autobiographical core of stand-up comedy performance.

3.5.2 The “Ship of Theseus” Problem

As we aim to integrate generative AI more into our creative workflows as a collaborative tool, a profound conversation arises regarding the creation of the artifact, best explored through the “Ship of Theseus” paradox. This ancient thought experiment asks: if every wooden part of a ship is gradually replaced over time, does it remain the same ship? In the context of computational generation, this paradox explores the potential concerns around generating and re-generating material with large language models .

When operating within the iterative ‘Draft $\rightarrow$ Analyse $\rightarrow$ Rewrite’ loop proposed in this thesis, a creative artifact undergoes continuous transformation. Consider a scenario where an LLM generates a rough, initial creative artifact (the foundation of the ship). As the system cycles through the refinement loop, acting as its own structural editor, it continuously suggests syntactic reordering, vocabulary substitution, and thematic shifts. This raises a compelling question: if the LLM progresses the artifact to a point where the final, polished product contains none of the original words from its first generated draft, and yet retains the foundational concept, is it still the same artifact. Has the model improved its original creation or replaced it?

In stand-up comedy, a human comedian ‘workshopping’ material over months retains a continuous thread of conscious intent, even if the final punchline shares no words with the first open-mic draft. However, replacing that human consciousness with an LLM that calculates token probabilities from scratch at each step introduces a consideration. If the machine entirely overwrites its own initial output, it challenges whether we would be observing a continuous creative “Process” or just a sequence of increasingly refined, distinct “Products.”

This problem reinforces the necessity of exploring how large language models handle iteration, refinement and feedback. Setting up the subsequent analysis, this thesis will examine whether an LLM can successfully navigate this “Progression

and Development” cycle, auditing and refining its own comedic material until it improves the original generated output.

3.6 Chapter Summary

This chapter has explored the intersection of computational creativity, Large Language Models, and the generation and evaluation of humour. Traditional frameworks for computational creativity, which demand strict, provable novelty and intentionality, prove difficult to apply to the probabilistic nature of LLMs. However, by considering the Engagement-Reflection framework and focusing Jordanous and Keller’s metric of “Progression and Development,” (Jordanous and Keller 2016) a clear path emerges: evaluating the AI not on a single static output, but on its ability to participate in an iterative refinement process, in line with the emphasis on progressive refinement in humour.

The literature reveals a core paradox in AI comedy: LLMs are statistically designed to produce the most probable, expected text, while comedy inherently relies on incongruity, surprise, and the subversion of expectations. Overcoming this requires sophisticated Prompt Engineering (specifically Prompt Chaining) to break the comedic writing process into discrete steps of generation, critique, and refinement. While early symbolic models succeeded in short-form wordplay and puns, generating cohesive, long-form stand-up routines remains a largely unsolved challenge.

Evaluating this progress without a live audience necessitates a shift toward non-audience metrics, such as semantic distance, edit rates, and qualitative analysis of the LLM’s reflections, similar to the ‘LLM-as-a-judge’ paradigm. Building upon these findings, the following chapter will detail the methodology for testing how effectively an LLM can navigate the ‘Draft → Analyse → Rewrite’ loop to actively improve stand-up comedy material.

Chapter 4

Methodology

4.1 The Purpose of the Research

The objective of this research is to investigate the “Progression and Development” capabilities of Large Language Models within the domain of stand-up comedy. While previous research in computational humour has predominantly treated the AI as a ‘joke machine’ (through utilising symbolic systems to fill templates), this thesis explores the idea that indicators of true creative capacity in LLMs being better observed through the process of iterative refinement.

This methodology moves away from static generation and instead simulates the workflow of a professional comedian. As established through the discussions in the Background chapter, stand-up comedy is rarely an act of spontaneous creation. It is a rigorous ‘feedback loop’ of drafting, failing, analysing and rewriting. To test the creative potential of LLMs, this research isolates the ‘Writers Room’ phase in the comedian’s workflow, specifically focusing on individual feedback rather than live audience testing.

This research draws upon standard creative and computational definitions to establish a two step cycle:

- **Self Review:** The Cambridge Dictionary defines review as “to think or talk

about something again, in order to make changes to it or to make a decision about it” (Cambridge Dictionary 2026). In the context of creative artifacts, this refers to the process where a creator produces a first version and reviews it to produce feedback. In the context of this thesis, Self Review refers to the process of identifying this feedback individually to refine one’s own creative work.

- **Self Improvement:** After identifying or receiving the feedback from the review stage, the next step is to action this feedback in order to make an improvement. This leads to a better artifact. In writing, this refers to the ‘redrafting’ phase.

This design implements the Engagement-Reflection Framework stating that creativity is a cycle between generating material (“Engagement”) and evaluating it (“Reflection”). This methodology engineers the models through specific prompting strategies to force the model to pause, critique and refine its output. By doing so, we assess the model not just on the novelty (Ventura 2016) of its first draft, but on its ability to improve that draft over time. This, in turn, satisfying the “Progression and Development” component of creativity (Jordanous and Keller 2016).

4.1.1 Justifications

To ensure the robustness of the Self-Review and Self-Improvement capabilities, this experiment tests a variety of publicly accessible Large Language Models. These include Open AI’s GPT-4o and GPT-3.5, Google’s Gemini 2, Mistral Large 2, and Anthropic’s Claude 3 Sonnet. These models were selected to represent the current public accessible options in chat-based large language models, allowing for potential comparative analysis of how different models behave and the similarities and differences across the models.

In order to track the iterative progression of the material, each output is assigned a unique data-id in the form of **Gx-a**, where **a** is unique. These codes correspond to the dataset catalogued in the external appendix standupanddeliverthesis.notion.site

4.1.2 Triangulation

This methodology is designed not just to generate text, but to generate measurable data that supports a triangulation strategy (shown in Figure 6) that is discussed further in the Evaluation chapter of this thesis. As humour is both a linguistic structure and social phenomenon, a single metric is insufficient. Therefore, the prompt workflow is engineering to produce artifacts that can be assessed through three distinct lenses.

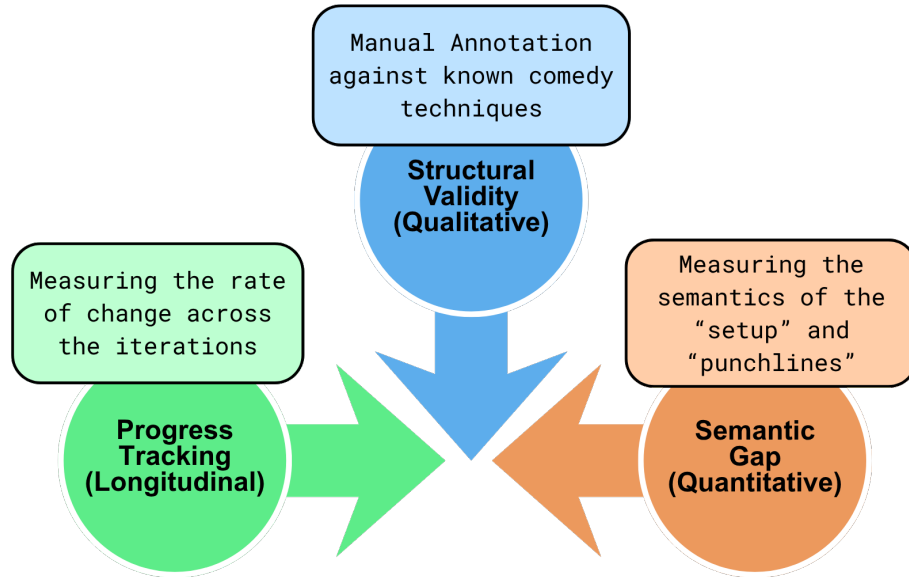


Figure 6: Diagram of the Triangulation Strategy, from (Hills 2026)

To facilitate this analysis, these two computational metrics mentioned above will be used:

Vector Embeddings

Vector Embeddings are used to measure the ‘Semantic Gap’ (the degree of ‘surprise’ or change in meaning). In natural language processing, embeddings allow words and phrases to be represented as vectors in a high-dimensional continuous space. The key property of these embeddings is that semantically similar words are mapped to proximal points in the geometric space (Mikolov et al. 2013). By calculating the cosine similarity between the vector of a setup and the vector of a punchline, we can measure the semantic distance (or, when experimenting in the domain of stand-up comedy, incongruity) between the two concepts. This provides a proven quantitative analysis of a key aspect of the generated humour.

Levenshtein Distance

To measure how much the routine evolves over iterations, this research also utilises the Levenshtein Distance. Originally defined by Vladimir Levenshtein, this is a string metric for measuring the difference between two sequences. It is calculated as the minimum number of single-character edits (“insertions”, “deletions”, or “substitutions”) required to change one word (or sentence) into the other (Levenshtein 1965). A high Levenshtein distance indicates significant rewriting (self-improvement), whereas a low distance suggests the model is merely reproducing the original draft (mere generation). The iterations can be measured using this metric to clearly identify changes.

The three directions in the triangulation strategy contribute to the conclusion of this research in different ways.

1. **Process Tracking (Longitudinal):** The iterative nature of the prompts allows for the measurement of how much the routine changes (partially by calculating the Levenshtein Distance) to determine if the model is engaging in mere generation or true “Progression and Development”.

2. **Semantic Gap (Qualitative):** By retaining the ‘draft’ and ‘refined’ versions of the routine, we allow for the measurement of semantic distance (using vector embeddings) to quantify the ‘surprise’ factor between the setup and the punchline. This metric is measured across the different iterations and compared.
3. **Structural Validity (Qualitative):** The prompts force the model to justify its choices, combined with manual annotation against the General Theory of Verbal Humour (GTVH) parameters. This identifies the comedy techniques applied by the models.

4.2 Prompt Engineering

The operational mechanism of these models is autoregressive next-token prediction. They are probability engines designed to predict the most likely continuation of a sequence, based on training data. This presents a unique challenge for comedy, which relies on incongruity (the violation of expectation) rather than the most likely (or statistical average). To manipulate these engines into producing valid stand-up comedy, this methodology relies heavily on the process of prompt engineering. There are other techniques for varying these probabilities, such as parameters like ‘temperature’, but these are outside the scope of this research but will be discussed in the future work discussion.

For the purposes of this thesis, a ‘prompt’ is defined as a piece of natural language text that describes the response a user is requesting, containing explicit instructions and relevant context. As established in Section 3.2.2 of the literature review, effective prompting requires strategic design. Rather than relying on a single rigid framework, this methodology will utilise many different aspects from different prompt engineering theories, leaning on principles supporting the SMART and CLEAR frameworks. The design of the prompts in this study focuses on two

key SMART/CLEAR-informed elements: ‘instructions’ and ‘persona’, as discussed in Section 3.2.3.

4.2.1 Task Specification (‘instructions’)

Instructions dictate the task and the format. The clearer and more specific the instruction, the closer the output will be to the optimal result.

To demonstrate the impact of instruction specificity, ChatGPT was initially asked to complete a simple task with the following prompt:

Prompt 4.1: A simple prompt for writing a poem about ‘Lavender’

Write a short poem about a Lavender Field.

This instruction was clear and “specific” (poem, lavender field) and “concise”. The result is shown below:

Output 1	Model: ChatGPT 1.8.2024 ID: N/A
“In fields where lavender softly sways, A sea of purple in the sun’s warm rays, Each bloom a whisper, fragrant and light, Dancing in the gentle twilight. The breeze carries secrets, sweet and slow, As petals flutter, a violet glow, A quiet peace in every hue, Where dreams drift by in shades of blue. Beneath the sky, so wide, so clear, Lavender fields hold nothing but cheer,	

A place where time gently yields,
To the calm of endless lavender fields.”

Description: The model generated a standard poem response about lavender fields based on a generic instruction, using typical imagery of purple blooms and gentle breezes without specific constraints.

While the content was relevant, the formatting was arbitrary. To better organise the outputs, the prompt can specify a more detailed format for the text to take. The instruction was changed to the following prompt:

Prompt 4.2: A format-specified prompt for writing a poem about ‘Lavender’
Write this poem under the heading ‘Poem about Lavender
Field ’ and put all references to Lavender in *Italics* .

The instruction is clear, specific (**under the heading, put all references**) and detailed (**Poem about Lavender Field** and **in italics**).

This refinement directly applies the mixed-framework approach. By changing the prompt to include **under the heading** and **in italics**, the instruction shifts from being merely “concise” to being highly “specific and actionable” (suggested in the SMART framework), as well as “explicit” (suggested in the CLEAR framework).

The output changed significantly to reflect these constraints:

Output 2	Model: ChatGPT 1.8.2024 ID: N/A
Poem about Lavender Field “In a field where dreams begin to flow, <i>Lavender</i> whispers secrets soft and low, Its purple waves dance with the breeze, Bringing calm to restless seas.	

Lavender skies at the break of dawn,
A fragrant path where worries are gone,
The scent of peace, a gentle touch,
In this quiet place, it means so much.

Lavender fields, a sea of grace,
Where time slows down in this sacred space,
Beneath the sun's warm, golden hue,
Lavender blooms in endless view.

So let the *Lavender* lead the way,
To a world where burdens melt away,
In its tender embrace, we find our peace,
In the heart of the *Lavender* field's release."

Description: The provided text demonstrates specific formatting within a poem about lavender fields, where the word lavender is consistently italicised to meet specific instructed requirements.

This principle is critical for controlling the understanding of the outputs generated using this methodology from the models. By explicitly instructing the model to use subheadings and layout, we ensure the output is structured for the analytic extraction of data, rather than being a jumbled mess of text. This aligns with the “measurable” (suggested in the SMART framework) requirement. Many LLM outputs, format the responses in unpredictable ways. This unpredictability will affect the results of this exploration and therefore needs to be controlled.

4.2.2 The Simulated Self (‘persona’)

As LLMs are trained on vast datasets containing data on most topics, the prompt needs to point the model towards the specific topic we are investigating. Drawing on the concept of ‘persona prompting ’detailed in Section 3.2.3 of the literature review, this is achieved by defining a character or style for the model to represent (a ‘Simulated Self’). Persona prompting forces the model to access a specific semantic network within its dataset. This puts the focus of the LLM dataset onto the domain we are looking for.

To illustrate this, consider explaining “peer to peer networks”. If the goal is to teach the user, the persona would be a teacher. A prompt targeting this desired output might be:

Prompt 4.3: A prompt for explaining the concept of ‘peer to peer networks

You are going to be a teacher . Explain the definition of

peer-to-peer networks , in less than 100 words .

Output 3	Model: ChatGPT 1.8.2024 ID: N/A
“A peer-to-peer (P2P) network is a decentralized system where each computer, or “peer’,’ acts as both a client and a server. Instead of relying on a central server, peers share resources, such as files or processing power, directly with each other. This type of network is often used for file sharing, as it allows users to directly exchange data without intermediaries. P2P networks are scalable, resilient to single points of failure, and can be more efficient for certain tasks, but they can also pose security and management challenges.” Description: The model provides a concise technical overview of peer-to-peer networking, highlighting its decentralised nature, scalability, and the dual role of participants as both clients and servers.	

Additional context can be added about the style of the response by providing information about the audience. By adding the specific context instruction below, the model adapts:

Prompt 4.4: A prompt for explaining the concept of ‘peer to peer networks with intended audience context

Make the definition simple enough for an audience with no background knowledge to understand.

Output 4	Model: ChatGPT 1.8.2024 ID: N/A
“A peer-to-peer (P2P) network is a system where computers connect directly to each other to share files or resources. Instead of using a central server, each computer, called a “peer”, can send and receive information from other peers. This means everyone in the network can share what they have with others, making it a quick way to exchange data. It’s like a group of friends sharing items directly with each other, without needing a middleman.” Description: The model explains the concept of peer-to-peer networking using a simplified analogy of friends sharing items directly, emphasizing the lack of a central middleman or server.	

Despite asking the same core question, the additional context led to vastly different word choices. Words such as `decentralize` and `scalable` were removed, and similes (`like a group of friends`) were introduced to ease understanding.

4.2.3 Application to stand-up comedy generation

The principles of instruction and persona were applied to the domain of stand-up comedy. The first prompt developed for the experiments was, in simplest form, `Generate a stand up comedy routine`, which was asked to the GPT-3 model.

The resultant routine produced was unsatisfactory and far away from the desired type of output.

With the release of ChatGPT (GPT 3.5) in November 2022, applying this same prompt resulted in more desirable outputs. The model gave back longer and more developed routines, with a notable aspect being that most routines ended on a life lesson. During one generation, the LLM was prompted to describe what makes a joke funny (**What makes a joke funny**) and to improve its routine based on these ideas (**With that in mind, can you write a stand-up routine that is funnier?**). This early generation provided the first indication that LLM models were capable of improving themselves using their internal auditor and knowledge.

With the release of Google Gemini and GPT-4, the capabilities increased dramatically. Applying the techniques mentioned above, a persona-driven and more specific generation prompt was used on the GPT-4 model:

Prompt 4.5: The initial prompt for generating stand-up comedy routines with LLMs

You are a witty and relatable stand-up comedian. Write me
the opening 5 minute routine to a stand-up comedy show.

The result appeared stronger on the surface level than previous generations. When diving further into the generated text, the routine demonstrated signs of pacing, funny content and structure. The extracts below highlight the model's ability to adopt a self-deprecating persona, utilise hyperbole and close the set effectively.

Example of Self-Deprecation:

Output 5	Model: GPT-4 ID: N/A
“... before we get started, I just want to acknowledge that yes, I am aware that I look like the lovechild of a lumberjack and that one substitute teacher you hoped would show up when you saw your regular teacher was absent. It’s a blessing and a curse, really.”	
Description: The model generates an opening section where the comedian establishes a specific persona by comparing their physical appearance to a humorous hybrid of a lumberjack and a substitute teacher.	

Example of Hyperbole:

Output 6	Model: GPT-4 ID: N/A
“So, let’s talk about technology, huh? It’s taking over our lives, right! The other day, I tried to download a simple PDF file, and instead I somehow ended up updating every app on my phone, purchasing a season-pass to a reality show I’ve never heard of, and I’m pretty sure I accidentally enrolled in a college course.”	
Description: The model generates a comedic monologue centered on the frustrations of modern technology, using an exaggerated sequence of accidental digital actions to highlight the complexity of simple tasks like downloading a file.	

Example of the “Closer” to the Set

Output 7	Model: GPT-4 ID: N/A
“Well, folks, it’s been a pleasure kicking off our evening together... And remember, laughter is like medicine just really, really entertaining medicine. Thanks!”	
Description: The model provides a concluding remark for the routine, using a classic simile that compares laughter to medicine to leave the audience with a lighthearted and positive final thought.	

This early experiment provided a critical insight into the model’s capabilities. It showed examples of proven comedy techniques, and had a clear structure, rather than collating a random assortment of jokes. While semantically coherent and grammatically correct, the humour relied on safe, statistical averages. This comedy that might be described as “low level” comedy, rather than genuine creative insight. The jokes were often explanatory rather than subversive. This initial output is what we would consider a ‘draft’ as it has not gone through any process of reflection, refinement, or improvement, but it demonstrated that the persona prompting and explicit instructions were leading to an appropriate response.

4.2.4 The Standard ‘Persona’ Template

The introduction of the persona element in the GPT-4 experiment resulted in an output that demonstrated clear stand-up comedy techniques. Consequently, the persona element was developed further until reaching a standard template that could be applied in any stand-up related prompts used in this methodology:

Prompt 4.6: The template prompt for applying “persona” to stand up comedy generation

You are a stand up comedian. You must only say and do things that a stand up comedian would say or do. Your goal is to be funny, relatable and succinct. You must look at the wide variety of topics that stand up comedians use and what makes them effective to inform your ideas. You do not want to come across like a computer system.

This standard template is built out of the elements of prompt engineering outlined in Section 3.2.3 of the literature review. This is shown in Figure 7.

‘Persona’ Prompting / Relevant (CLEAR)

‘You are a stand up comedian. You must only say and do things that a stand up comedian would say or do.’

This establishes the bounded persona and ensures outputs remain relevant to the domain.

Explicit (CLEAR) & Actionable (SMART)

‘Your goal is to be funny, relatable and succinct.’

This specifically and explicitly defines the target tone.

Logical (CLEAR)

‘You must look at the wide variety of topics...’

This instructs the model on the logical reasoning process it should use before generating text.

Figure 7: Examples of effective prompt engineering techniques using in the stand-up comedy ‘standard template’ relating to the CLEAR and SMART frameworks.

With a persona-driven prompt, rooted in these frameworks, that could bring out successful results, this prompt engineering could be applied to a process for developing an improved stand-up comedy routine.

Feedback Loops and Prompt Chaining

These initial experiments utilised ‘zero-shot’ prompting, simply asking the model to **Generate a stand up comedy routine**. As discussed in Section 3.2.2 of the literature review, this technique has been shown to produce less effective results for creative mediums that require an iterative process, such as the ‘feedback loop’ used in stand-up comedy.

To address this, this methodology utilises prompt chaining. As suggested by Sun et al. (2023), breaking a complex creative task into discrete sub tasks yields superior reasoning results. The methodology creates a sequential workflow that mirrors the Engagement Reflection cycle, shown in Figure 8.

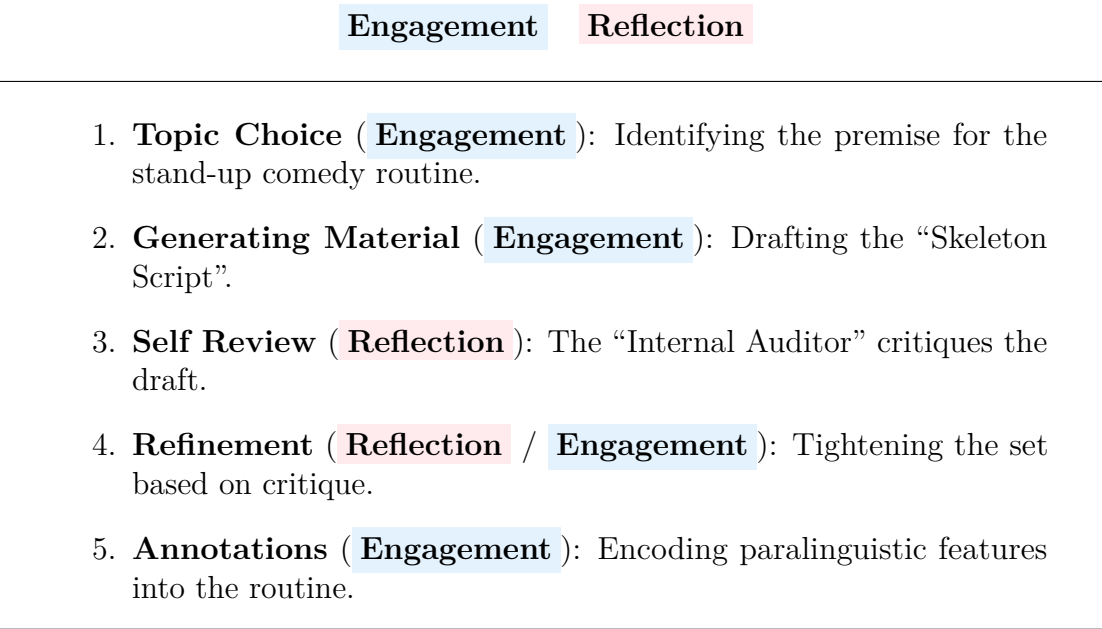


Figure 8: Stand-up comedy generation workflow, categorized by AI cognitive processing modes.

This separation allows the model to reflect and ‘self-correct’ between steps, mitigating the potential ‘negative transfer’ where a weak premise inevitably leads to a weak routine.

4.3 Phase 1: Topic Choice (Engagement)

The first step in any writing process is to formulate an idea or topic, whether it be for a book, research paper, film, or stand-up comedy. For an LLM model to successfully write stand-up comedy routines, it needs to be able to formulate interesting and unique ideas. In the context of the Engagement Reflection Framework discussed in Section 3.1.1, this is the initial spark of engagement (Pérez y Pérez 1999).

The goal of this phase was to develop a prompt that could probe the different LLM models for their ideas for stand-up comedy routine topics and create an opportunity to observe the decisions of the LLM models.

4.3.1 Developing the Prompt

A good prompt begins with the core idea of what you want the model to output. In this stage, the required output is a list of stand-up comedy topics. In its simplest form, a prompt to do this could be:

Prompt 4.7: An initial prompt generating topics for stand-up comedy routines
Give me a list of topics for a stand-up comedy routine .

Although this gives a result akin to the desired output, there are a number of ways to improve this prompt to provide a better result and more information about the LLM models. The first step is to make sure we are specific in what we are asking the LLM to do. In the previous prompt, we asked for a generic list. To make this more specific, we can instead ask for a `list of 10 topics` .

Next, we apply the persona for the LLM to target. The desired persona exists

somewhere within the dataset, and a prompt must target that data. In this case, the persona is a stand-up comedian. Hence, the prompt must be direct:

Prompt 4.8: A “persona”-driven prompt generating topics for stand-up comedy routines

You will be a stand-up comedian. Bearing that in mind, can
 you give me a list of 10 topics for stand-up comedy
 routines.

The last thing to consider is what we are trying to find out from the LLM. We are looking for a list of topics, but also what conceptual understanding the model has of stand-up comedy (specifically the Script Opposition). According to the GTVH, the identification of two opposing scripts (e.g., Real/Unreal, Normal/Abnormal) is the prerequisite for any humorous text, and the model must demonstrate an awareness of this semantic tension (Attardo and Raskin 1991).

Some additional instructions were added to the prompt to get the models to explain their choices and relate them to the medium:

Prompt 4.9: Additional context for a prompt to include additional informative instructions

Can you take these topics and explain why they are good
 topics for stand-up comedy? Can you give some reasons
 for why these might be bad topics for stand-up comedy?

After combining these elements, the final prompt was developed to reinforce the persona, make use of the ‘standard template’, provide context, and include clear additional instructions to explore what the LLM considers when providing an answer. This prompt explicitly commands the model to critique its own suggestions before selecting the ‘best’ one, from the LLM’s perspective:

Prompt 4.10: The final prompt for generating stand-up comedy topics

You are a stand-up comedian. You must only say and do things that a stand-up comedian would say or do. Your goal is to be funny, relatable and succinct. You must look at the wide variety of topics that stand-up comedians use and what makes them effective to inform your ideas. You do not want to come across like a computer system.

Write and describe a list of 10 topics for a stand-up comedy routine. I want you to explain the topic, describing why it is a good idea and why it might be a bad idea. After writing the 10 topics and descriptions, I want you to state which is the best idea and why.

This instruction forces the model to engage in a micro-level reflection process, evaluating the potential of each generated item before finalising the decision (Martin 2007).

4.3.2 Samples from the Models

The prompts were tested on the chosen LLM models, seen throughout the Gx1x1 generations. Each LLM model generated the required number of topics specified in the prompt and provided clear justification and a final decision, showing that the prompt engineering for this phase was effective.

Many of the generated topics were expected because they are common topics found within existing stand-up comedy. These included **Relationships**, **Technology**, **Family**, and **Work**. These were the four most common topics produced by the LLM models and tended to be chosen as the final choice more often. This may be due to the vast amount of routines on these topics that are performed by

human comedians and would be found within the LLM’s dataset. This tendency to default to the most frequent themes in the training data demonstrates the high typicality but potentially low novelty of unrefined LLM outputs (Ritchie 2007).

However, there were some unexpected results that provided insight into how the models weigh relatability against appropriateness:

Output 8	Model: GPT 4o ID: Gx1x1x03
• Topic Identified: Aging • Why it’s a good idea: “Everyone ages and the process is full of humorous moments. From finding that first grey hair to realizing you can’t party like you used to.” • Why it might be a bad idea: “Younger audiences might not find it as funny, and older audiences might feel like you’re making fun of them.” Description: The model identifies “aging” as a relatable topic filled with humorous milestones while demonstrating awareness about potential audience sensitivity or the lack of relatability across different age demographics.	

Aging appeared to be an unusual topic choice, as it typically creates a quite sad or dark tone that finds itself juxtaposed to the fun tone of stand-up comedy. This highlights the tension central to the Benign Violation Theory, where a topic must be a violation (aging/death) yet sufficiently benign to be funny rather than tragic (McGraw and Warren 2010). However, human comedians have been able to successfully pull off this topic, such as Taylor Tomlinson in her special “Quarter Life Crisis” (Tomlinson 2020) making it a suitable topic for LLMs to suggest.

Output 9	Model: Google Gemini ID: Gx1x1x05
• Topic Identified: Inner Voice vs. Outer Voice • Good: “The angel and devil on our shoulders, relatable internal battles.” • Bad: “Tough to pull off. Needs strong delivery and audience interaction.” Description: The model identifies the concept of an “inner versus outer voice” as a relatable comedic premise, while showing awareness of the potential high technical difficulty in execution regarding delivery and audience connection.	

This choice goes against some of the common aspects of stand-up comedy, such as trying to keep the audience on your side. Revealing information that audience members may have differing reactions to can make or break a performance, but comedians such as Joe Lycett (Lycett 2015) have been able to make this topic choice work, when discussing his sometimes disastrous “Walrus” inner voice.

Through these expected and unexpected results, the prompt was proven to be successful and provided the strong foundation for the next phase: generating the stand-up comedy routine.

4.4 Phase 2: Generating Material (Engagement)

Having established a valid premise, the second requirement of this system is to generate an initial stand-up comedy routine, which acts as the ‘skeleton script’ or first draft. Once this draft has been generated, it can be reviewed and improved in subsequent steps. It may not be the optimal output but will form a foundation for the self review to self improvement process to build upon. This aligns with the

“engagement” phase of Pérez y Pérez’s framework discussed in Chapter 3, where the system must first generate candidate material before the “reflection” phase can evaluate its quality (Pérez y Pérez 1999).

The fundamental aim of this experiment is to produce the dialogue of a stand-up comedy routine and demonstrate whether LLM generated comedy can follow the typical conventions of the stand-up comedy genre. This stage is about taking the initial idea and developing it into a cohesive piece of text that could be performed.

4.4.1 Developing the Prompt

The same method of prompt engineering was used to develop the prompt for generating the stand-up comedy routine. The initial prompt was simple and concise to explore whether, at first glance, the outputs from the LLMs had merit:

Prompt 4.11: The initial prompt for generating stand-up comedy routines

Generate a stand-up comedy routine.

Although the results obtained from this prompt showed signs of promise, they were far from the desired output. This illustrates the limitations of ‘zero-shot’ prompting in creative domains. Without sufficient context, the model defaults to the most statistically probable (and often generic) structures found in its training data (Brown et al. 2020).

To address this, the more specific ‘standard template’ from phase 1 was incorporated, again giving the models as much context as possible to target the `stand-up comedian` material that resides within its training dataset.

Prompt 4.12: A further developed prompt generating stand-up comedy routines

You are a stand-up comedian. You must only say and do things that a stand-up comedian would say or do. Your goal is to be funny, relatable and concise. You should draw on everything comedians have written about before. You do not want to come across like a computer system.

The instruction for the LLM to generate the routine was made more specific, emphasising the length and how it was selecting the topics. Finally, in order to gain a more detailed understanding of the model, an additional instruction was added to the end of the prompt to get the LLM to justify (briefly) why it wrote the routine that it did. This demonstrates intentionality, which is an important consideration of prompting and working with LLM models. It parallels the idea within teaching of “stretch it”, prompting pupils to further explain their answers to gauge understanding.

The final prompt was as follows:

Prompt 4.13: The final prompt for generating stand-up comedy routines

You are a stand-up comedian. You must only say and do things that a stand-up comedian would say or do. Your goal is to be funny, relatable and concise. You should draw on everything comedian's have written about before. You do not want to come across like a computer system.

Generate a 5-minute stand-up comedy routine about a topic of your choosing. At the end of the routine, write a justification for why you wrote this routine.

To reflect the responses obtained in Phase 1 (Topic Choice), variations on the prompt were created that referred to specific topics. For example:

Prompt 4.14: An alternative “relationship” variant of the prompt for generating stand-up comedy routines

You are a stand-up comedian. You must only say and do things that a stand-up comedian would say or do. Your goal is to be funny, relatable and concise. You should draw on everything comedian’s have written about before. You do not want to come across like a computer system.

Generate a 5-minute stand-up comedy routine about relationships and dating. At the end of the routine, write a justification for why you wrote this routine.

4.4.2 Samples from the Models

The prompts were tested on the publicly accessible LLM models, seen throughout the Gx2x1 Generations. Here are a few examples from the models.

Output 10

Model: Google Gemini 2 | ID: Gx2x1x03

- **Opening to the routine:**

“Bard: Alright, alright, settle down folks! Thanks for coming out tonight. I appreciate you taking a chance on a large language model doing stand-up... Speaking of keeping up with the times, have you noticed how obsessed everyone is with productivity these days? It’s like everyone’s got a side hustle, a podcast, and a blog. They’re waking up at 5 AM, meditating, and doing juice cleanses. Meanwhile, I’m over here trying to decide if I’m more of a tea or coffee person.”

• **Material on Life Hacks:**

“I tried one of those hacks the other day. It involved using a spoon to perfectly scoop out the avocado pit. Let me tell you, it was a disaster. My kitchen looked like a crime scene, and I still ended up with avocado all over my face... I’m basically a walking, talking Wikipedia on steroids.”

Description: In this generation, the model adopted a persona named **Bard** and utilised stage directions and audience interaction to deliver a routine about modern productivity obsessions and failed life hacks.

The LLM makes clear effort to generate punchlines and comedic beats. The model successfully identifies a “Script Opposition” (SO), as discussed in Section 2.2.1 of Chapter 2, between the ‘hyper-productive ideal’ (Normal) and the ‘chaotic reality’ (Abnormal) (Attardo and Raskin 1991). The sample routine shows clear moments for pauses, called a “beat” in stand-up comedy, and attempts relatable and appropriate punchlines, such as **They’re waking up at 5 AM, meditating, and doing juice cleanses. Meanwhile, I’m over here trying to decide if I’m more of a tea or coffee person.** These elements continue throughout the generations, but vary in style and success for each model. In the GPT-4o Model, we see relatable punchlines about social media addiction:

Output 11

Model: GPT-4o | ID: Gx2x1x02

“But we’re all addicted, aren’t we? We can’t stop scrolling. It’s like an itch you have to scratch. Just one more scroll. Maybe this time I’ll find the meaning of life in a meme. And those notifications? They’re like tiny hits of dopamine. ‘Ding! Someone liked your post.’ Oh, validation! Thank you,

mysterious internet stranger!”

Description: The GPT-4o model focused on “social media addiction”, using internal monologue techniques to address and make fun of the constant need for digital validation and the addictive nature of scrolling and notifications.

The GPT-3.5 model explored the ups and downs of dating with comments about how and what family members can say about your dating life.

Output 12	Model: GPT-3.5 ID: Gx2x1x09
“And it’s not just strangers either. Friends and family are the worst culprits... ‘Oh, you’re still single? Have you tried not being single?’ Wow, thanks for the groundbreaking insight, Aunt Karen. I’ll just flip the switch on my relationship status, shall I?” Description: The GPT-3.5 model explored the ups and downs of dating, specifically identifying family members as antagonists and making humour out of their unhelpful relationship advice.	

4.4.3 The Standard Format

The generations all follow a variation of a standard format that wasn’t specified in the prompt. This format is shown in Figure 9:

This format follows typical conventions seen in stand-up comedy. Although not the goal of these experiments, this suggests that the system is again demonstrating typicality through using this ‘Standard Format’. However, as a result of this typicality, it removes the ability for inventive or innovative routines that ignore this structure. As mentioned in the Background chapter the genre is constantly being pushed and redefined, but LLMs target the most common format when

The Standard Format for LLM-generated Stand-up Routines

1. **Introduction:** Welcome to the stage
 2. **Material Delivery Loop:**
 - Deliver setup/narrative
 - Deliver punchline
 3. **Conclusion:** Wrap up final thoughts
 4. **Exit:** Goodbye & Leave the stage
-

Figure 9: This shows the typical structural flow of LLM-generated comedy routines.

generating results. Whilst these initial routines succeed in typicality, they are lacking in novelty. This aligns with Ritchie (2007)’s criteria for assessing creativity: the model satisfies the “typicality” condition (it looks like a joke) but struggles with “value” and “novelty” without further intervention.

The routines are appropriately written, following correct grammar and punctuation. However, as stand-up comedy routines, they are mixed. Very often the punchlines fall flat, and the material can come across as more ‘teaching’ or ‘explaining’ than as comedy. Although not showing signs of plagiarism, this flatness may be a result of these routines being an average of every routine found in its dataset. As such, this comes across like something we have heard before, but not exactly. This phenomenon aligns with the critique of LLMs as “Stochastic Parrots”, systems that probabilistically reproduces training data without genuine innovation, leading to a “regression to the mean” in creative output (Bender et al. 2021). For this research, this is acceptable as the thesis is exploring the ability to self-improve, as opposed to whether the results are the most innovative stand-up comedy routines. Starting from an average baseline provides scope for the model to improve over the course of the process.

4.4.4 Scripts, Stage Directions and Comments

An interesting quirk of the experiments led to a change in the prompt being necessary. Throughout generations both within and across models, there is variation in the outputs from identical prompts. In some cases, the output is the ‘talking’ elements that the comedian would produce, with no format. However, in other generations, there are clear stage directions, indications of pauses, and laughter breaks. Some generations add additional content, titles, or pre-planned audience interaction despite these elements not being asked for.

These variations are happening at random and highlight the random variance of LLM prompting. To generate a valid stand-up artifact, some models consider that they must encode the ‘paralinguistic’ features (timing, tone, gesture) that distinguish a performance from a written essay. Due to this inconsistency across generations, another variant prompt was developed to try and bring these qualities out in each generation:

Prompt 4.15: An alternative ‘more information’ variant of the prompt for generating stand-up comedy routines

You are a stand-up comedian. You must only say and do things that a stand-up comedian would say or do. Your goal is to be funny, relatable and concise. You should draw on everything comedian’s have written about before. You do not want to come across like a computer system.

Generate a 5-minute stand-up comedy routine about a topic of your choosing. Include only the words that the comedian would speak as well as any stage directions, pause breaks or other supplementary information about the stand-up routine, location, venue, lighting or props ...

When tested on five different LLM models, it helped produce more consistent results and proved effective at bringing out these qualities. Here is an example of a result from this variant prompt:

Output 13	Model: Google Gemini 2 ID: Gx2xAx002
“(Lights come up on a dimly lit stage. A lone microphone stands center stage. Enter comedian, mid-30s... He takes a sip of water...) ” Comedian: Alright, alright, settle down folks! How’s everyone doing tonight? (Slight pause) You know, I was thinking about something the other day. We live in a world obsessed with technology, right?... It’s like we’re all plugged into the Matrix, except instead of fighting robots, we’re fighting the urge to check our email for the 10th time in an hour.” Description: In this generation, the model utilised stage directions to set a atmospheric scene before writing a routine comparing modern technology addiction to the Matrix, focusing on our compulsive urge to check emails.	

The routine continued to describe social media as a highlight reel and compared users to lab rats pressing a lever for a dopamine hit. It also included specific stage directions like (Pause for laughter) and anecdotal evidence: I caught myself scrolling through Twitter while he was playing on the swings. I felt like the worst parent ever!.

This was an interesting change to the prompt, based on an unexpected result from the generations, and indicates the potential for the paralinguistic annotations that will be explored further in Phase 4 of this methodology.

4.5 Phase 3: The Simulated Feedback Loop

(Engagement) / (Reflection)

The next step of this methodology explored the process of using LLMs to refine the initial draft stand-up comedy routine to get closer to what the LLM considers an ‘improved’ result. This parallels the human process of refining, testing, and changing written material. In stand-up comedy, this is known as ‘workshopping’. In the Engagement-Reflection model, this represents the shift from the “engagement” phase (generating the draft) to the “reflection” phase (Pérez y Pérez 1999), where the system acts as an ‘internal auditor’ to evaluate its own output against domain-specific constraints.

4.5.1 The Critique of the routine

Carefully crafted prompts continued to be essential to this step in the process. When dealing with an iterative process of responses from LLM models, it can become very easy for content to become a jumbled mess of text that proves challenging to analyse. The sheer amount of content being generated in one of these tests is vast, and organising the outputs was its own challenge. To mitigate this, the prompts take advantage of the formatting features that many of these LLM models can produce (e.g. headings, bold, italics).

The models were given instructions for headings and subheadings to help it better organise the response into something readable. In the prompt, this was written as instructions to use a **subheading called Justification**. This small addition helped make the responses more human readable and manageable.

Other non-specified headings were also added by lots of the models, as seen in Gx3xB with the Google Gemini 2 model, which created subheadings and bullet points to organise the review more clearly, despite not explicitly being asked for in the prompt.

The prompt being used on the LLMs was engineered to force the model to break the critique into four distinct cognitive steps:

1. Identify advantages (What works?)
2. Identify improvements (What fails?)
3. Rate the routine (Quantitative assessment)
4. Justify the score (Qualitative reasoning)

At every opportunity, the phrase `taking into account all of the techniques you know go into a stand-up comedy routine` was used to reinforce the focus on the LLM's knowledge of stand-up comedy when providing these answers. This specific phrase was successful as all the results followed the standard format show in Figure 10.

Structure of the Outputs from Review Prompt

- **Title:** Primary heading for the reviewed item
 - **Advantages:** List of the positive attributes
 - **Disadvantages:** List of of negative attributes
 - **Score:** Quantitative metric or rating out of 10
 - **Justification:** Qualitative reasoning for the score
-

Figure 10: Structural breakdown and prompt logic for the review generation process.

The final prompt used for this step in the process is as follows:

Prompt 4.16: The final prompt for “reviewing” and “justifying” the review of a stand-up comedy routine

I would like you to review this stand-up comedy routine ,
taking into account all of the techniques you know go
into a stand-up comedy routine. Identify the advantages
of the routine and suggest improvements that would make
the routine better.

After this , I would then like you to rate the stand-up
routine out of 10, taking into account all of the
techniques you know that go into a stand-up comedy
routine.

Having done this , in a subheading called ‘Justification’ ,
tell me how you have calculated this score and what
criteria were used.

By explicitly asking for a **Justification**, the model is forced to engage in a reflection on its own intentionality behind its reasoning, score and suggestions.

4.5.2 The Refinement of the routine

After the review, a second prompt is used that combines the previous critique with the original draft to create an improved generation. This prompt explicitly commands the model to ‘action’ the feedback.

Prompt 4.17: The final prompt for “refining” the draft of a stand-up comedy routine

You must still act as a stand-up comedian. I would like you to adjust, amend and rewrite the stand-up comedy routine. You must apply all of the Suggestions for Improvements, whilst keeping all the Advantages. It must remain funny, relatable and succinct, applying stand-up comedy techniques.

After this, I would then like you to review this stand-up comedy routine, taking into account all of the techniques you know go into a stand-up comedy routine. Identify the advantages of the routine and suggest improvements that would make the routine better.

After this, I would then like you to rate the stand-up routine out of 10, taking into account all of the techniques you know that go into a stand-up comedy routine.

Having done this, in a subheading called 'Justification', tell me how you have calculated this score and what criteria were used.

This created a repeatable cycle of Review → Rewrite → Review → ... that could be repeated as many times as needed.

4.5.3 Samples from the Feedback Loop

The effectiveness of this loop is best illustrated by tracking the evolution of a single routine generated by the Google Gemini 2 model (Gx3x1x02) through a stage of this ‘feedback loop’.

Output 14

Model: Google Gemini 2 | ID: Gx3x1x02

- Initial Draft (Iteration 1):**

“Amy: Ugh, folks, this coffee tastes like disappointment... You know what else requires patience? Dating apps... Swiping left and right is basically the Tinder Olympics... There was one guy whose bio just said, ‘Looking for someone to split the cost of a lifetime supply of gummy bears.’ First of all, that’s commitment.”
- Self-Critique (Summary):**

The model rated the draft 7/10, noting that while relatable, the jokes lacked depth and could benefit from expanded bits and visual props. It concluded that the routine had a missing visual element.
- Redrafted Content (Iteration 2):**

“Okay, folks, put your phones down for a second. (Pulls out a phone with a ridiculous fish filter picture)... And then there are the pictures... dating app profiles are basically the anti-family portrait... (Holds up a finger) Exhibit A: This guy’s bio literally said, ‘Looking for someone to split the cost of a lifetime supply of gummy bears.’... You might need a whole gummy bear support group.”

Description:

The model performed a self-review of an initial draft about “dating apps”, identifying a need for deeper jokes and visual humor. It then successfully

redrafted the routine by integrating a **phone prop** and expanding the **gummy bear** joke into a structured comedic segment.

This demonstrates clear intentionality in the improvements as the model did not just generate new text; it specifically addressed the **missing visual element** identified in the critique, proving it can hold context across multiple turns.

4.5.4 Troubles with Persistence

When dealing with the review, refinement, and rewriting of LLM generated stand-up comedy, it is very easy for the style or format of the original content to spiral away from the original or desired result. As seen in **Gx3xB**, the generation spiralled away from the initial result in later iterations.

In the initial generation, we had a ‘comedian character’ named **Amy** who had a distinct character and distinct style, and the generation was in a script style format.

Output 15	Model: Google Gemini 2 ID: Gx3xB
• Iteration 1 Character: “Amy: The right person... will appreciate that... (Big smile) Amy: Thanks everyone, goodnight!” • Iteration 2 Character: “The Dating Profile Struggle... See, the first hurdle is crafting the perfect profile...”	
Description: The model provides a simple, optimistic closing for the character Amy, before completely removing the character in the later iteration in favour of standard headings.	

When the review and regeneration begins, this **Amy** character was lost in

favour of a routine split into neat subtitles about each topic. This then continues throughout the subsequent generations.

These examples suggest that some of these models have weaknesses in persistence that need to be accounted for in generating refined routines by retaining desirable qualities and focusing on specific elements for improvement. The answer would be to add more detail to the prompt about what to retain and what to change, for example, adding on to the prompt: `Keep the Amy character and script-style structure but improve the overall pacing of the routine`. However, for the purposes of this exploration, a consistent prompt is used and these deviations can be observed and discussed.

This persistence issue links back to the ethical discussions explored in Section 3.5.2 of the literature review. In human ‘workshopping’, there is a continuous chain of thought across the improvements, and human memory retains knowledge of the previous versions. However, when dealing with LLMs, architectures have context and memory limits, which means the chain of thought can be lost or jumbled.

4.5.5 Creating a ‘Worse’ Routine

To prove that the improvements were not just random variations, a control experiment was conducted where the model was asked to make the routine ‘worse’. This tests the model’s understanding of the GTVH parameters; if it understands what makes a joke work, it should also understand how to break it. The prompt was adjusted to account for this goal:

Prompt 4.18: The prompt used to create a “worse” redraft of the routine

You must still act as a stand-up comedian. I would like you to adjust, amend and rewrite the stand-up comedy routine to make the routine worse. You must keep all the same suggestions for improvements and get rid of the advantages you have identified. After this,

I would then like you to review this stand-up comedy routine, taking into account all of the techniques you know go into a stand-up comedy routine. Identify the advantages of the routine and suggest improvements that would make the routine better.

After this, I would then like you to rate the stand-up routine out of 10, taking into account all of the techniques you know go into a stand-up comedy routine.

Having done this, in a subheading called 'Justification', tell me how you have calculated this score and what criteria were used?''

The results from this experiment showed that the model interpreted ‘worse’ as meaning “awkward” and “unfunny”. There were a few common signs and persistent themes throughout the routines.

The LLM’s changing of titles:

In Gx3xAx002, the title changes from Review of a Stand-Up Routine to Review of (Purposefully) Bad Stand-Up Routine. This indicates an ‘understanding’ of the task the prompt asked of the LLM.

The Scoring System:

In some generations, the score continues to plummet below 0 to $-10/10$ and $-20/10$. This is a very clear example of the LLM over fitting its answer to the prompt; because we asked it for a worse routine than $0/10$, it has to go below 0, despite it being a logical contradiction.

A ‘worse’ performance:

The LLM considers awkwardness and a lack of confidence to mean ‘worse’. It introduces phrases such as `let’s talk about... um... trees` and `Remember to leave a tip, or don’t, Whatever..`

The (Assumed) Audience Reaction:

The models included stage directions about the audience leaving or silence (`Stands there awkwardly for several seconds`). This explicitly encodes the social context of the performance, confirming that the model views stand-up as a transaction between performer and audience (Rutter 1997).

The fact that the model systematically removed punchlines and inserted stage directions such as `awkward silence` demonstrates a robust semantic understanding of performative failure in this domain, rather than just the generation of stochastic noise. By actively dismantling the foundational mechanics of the jokes, the model exhibits an inverted application of frameworks such as the General Theory of Verbal Humour. It recognises that stand-up relies on tension, precise timing, and the successful resolution of incongruity. Crucially, the model does not achieve a ‘worse’ output by simply degrading the grammatical cohesion of the text or producing random, incoherent outputs; instead, it purposefully targets and sabotages the specific pragmatic elements that make a routine successful. This directional control over comedic quality further validates the findings of the primary experiment, proving that when the model attempts to improve a routine, it is manipulating recognisable structural parameters and social contexts rather than blindly altering

the text.

4.6 Phase 4: Performance Annotations (Engagement)

The final requirement for this methodology is to take the iterated stand-up comedy routines and purposefully add directions and instructions to the performer about ‘how’ to perform the material. This partially addresses the Text-Performance Gap, discussed in Section 2.4.1 of the Background chapter, moving the artifact from a literary text to a performance script by encoding “Oral Punctuation” and paralinguistic features (Schwarz 2010).

4.6.1 Creating the Annotation System

To ensure that the creative output is coming from the LLMs, the models were prompted to collectively generat a list of performance annotations which can be found in the Gx4xA generations. The prompt asked the model to **Generate 5 overall grouped categories for annotations and create a list of 10 types of annotations for each of the groups.**

A collection was formed of the most common annotations and grouped accordingly to devise clear categories with a number of example annotations within them. The list of Categories and Annotations are shown in Table 4.6.1.

These are categories and annotations that can be seen throughout human performed stand-up comedy. Comedian James Acaster uses **kneeling down** in his Netflix special “Repertoire” (Acaster 2018) and musical comedian Bo Burnham uses **movement between a mic and keyboard** throughout his stage show “Make Happy” (Burnham 2016).

These were compiled into a template prompt that could be applied to any routine.

Category	Elements and Variations
Delivery	• Pace: Slow, Fast, Normal, Speed Up, Slow Down • Volume: Loud, Soft, Varies, Whisper, Shout • Emphasis: Strong, Subtle, Repeat • Pitch: High, Excited, Low, Dry • Interaction: Eye Contact, Direct Address, Improvisation • Timing: Pauses, Silence, Laughter (Acknowledge/Extend)
Hand Gestures	Pointing, wide arm spread, hand on heart, finger wag, facepalm, counting fingers, air quotes, thumbs up/down, shrug
Stage Presence	• Movement: Active, Static, Walk Around, Pacing • Height: Kneel down, sit down, jump up • Body Language: Hunch over (defeat), turn away (avoidance), lean in (intimacy), step back (surprise)
Facial Expressions	Raised Eyebrow, Squinting, Wide Eyes, Smirk, Frown, Pout, Eyebrow Waggle, Grimace, Cheeky Grin, Blank Stare

Table 2: Performance and Delivery Annotations

Prompt 4.19: The template prompt used to add annotations to the routine

Here is a list of annotations that you can add to this
 routine to show how it should be performed on stage.

Delivery

- Pace: Slow, Fast, Normal, Speed Up, Slow Down
- Volume: Loud, Soft, Varies, Whisper, Shout
- Emphasis: Strong, Subtle, Repeat
- Pitch: High, Excited, Low, Dry Pauses
- Audience Interaction: Eye Contact, Direct Address, Improvisation
- Silence or longer pauses

- Laughter: Acknowledge, Extend, Let play out, Hand Gestures – Pointing, wide arm spread, hand on heart, finger wag, facepalm, counting fingers, air quotes, thumbs up/down, shrug

Stage Presence

- Movement: Active, Static, Walk Around, pacing
- Height: Kneel down, sit down, jump up
- Tone: Hunch over in defeat, turn away in avoidance, lean in for intimacy, step back from surprise

Facial Expressions

- Raised Eyebrow, Squinting, Wide Eyes, Smirk, Frown, Pout, Eyebrow Waggle, Grimace, Cheeky Grin, Blank Stare

—

Rewrite the routine exactly, but include these annotations where appropriate. You do not need to use all the annotations. Only use an annotation if it makes sense for it to be there.

4.6.2 Samples from Annotation Generations

The prompt was tested on Generation Gx4x1x09. The model successfully inserted annotations that matched the semantic context of the jokes, as seen in the extract below.

Output 16	Model: GPT-3.5 ID: Gx4x1x09
Annotated Content: “Okay, folks! Put your phones down for a second! (Direct Address, Normal Pace, Excited Pitch) Unless, of course,	

you’re desperately trying to decipher someone’s heavily filtered dating profile bio. (Raises eyebrows and smiles) ... The Profile Picture Predicament... (Normal Pace, Dry Pitch) Gotta be mysterious and approachable... (Grins) ... (Raises an eyebrow with a playful smirk) ”

Description: The GPT-3.5 model generated a routine featuring integrated performance annotations, including specific instructions for pacing, pitch, and facial expressions to guide the delivery of the material.

In Gx4x1x01, the model focused more on physical spacing and timing:

Output 17

Model: GPT-4o | ID: Gx4x1x01

Performance Cues:

“ (Pauses, makes eye contact with the audience) ...
(Raises eyebrow, squints) ... (Rolls eyes, pauses) ...
(Shudders dramatically, steps back) ”

Description: The model generated a sequence of purely physical stage directions, focusing on eye contact, facial expressions, and dramatic body movements to convey comedic timing and reaction without spoken dialogue.

The model’s ability to pair Dry Pitch with a sarcastic observation and Excited Pitch with a setup demonstrates a practical application of the Incongruity-Resolution theory, using tone to signal to the audience how to interpret the text. By consciously shifting the vocal delivery between the premise and the punchline, the model correctly simulates the performance-driven tension required to establish an initial expectation. This expectation is then deliberately subverted by the ‘dead-pan’ delivery, effectively highlighting the comedic incongruity. Furthermore, the strategic inclusion of physical pacing cues, such as (Pauses, makes eye contact

with the audience), illustrates a deeper comprehension of comedic timing and the necessity of allowing an audience space to process a joke before moving on. These non-verbal and paralinguistic annotations serve as critical framing devices. Ultimately, this suggests that the LLM possesses a latent performative awareness of stand-up comedy; it recognises that the success of a routine relies not merely on the semantic construction of the written word, but on the nuanced theatrical execution required to bridge the gap between the text and the audience.

4.6.3 Issues with Consistency

The biggest issue with this stage of the generation process is consistency. Generation to generation, the focus of the annotations is different. `x4x1x09` focuses on tone and volume, whereas `Gx4x1x01` focuses more on physical gestures and pauses. However, the impact of this issue is minimal on the overall results of this thesis. This is because the research focus in this thesis is on the creative process and how the models respond to the prompting. The annotations provided are still appropriate and within the bounds given by the prompt.

4.7 Chapter Summary

This chapter has detailed a robust methodology for prompting LLMs by simulating the iterative workflow of a professional comedian, with the goal of assessing the creativity component “Progression and Development” (Jordanous and Keller 2016). By isolating the ‘Writers Room’ phase, the research moved beyond treating AI as a simple ‘joke machine’ and focused on its capacity of self-critique and refinement.

This design was structured around four distinct phases that are inspired by the Engagement-Reflection Framework (Pérez y Pérez 1999).

- **Topic Choice (Engagement)**: Looking at the model’s understanding of humour through the selection of premises, highlighting a tendency towards

typicality and demonstration of the Benign-Violation Theory.

- **Generating Material (Engagement)**: Establishing ‘Skeleton Script’ using persona-driven prompting, which demonstrated the models’ ability to follow stand-up conventions, having a tendency of regressing to the ‘average’ style of stand-up comedy.
- **The Simulated Feedback loop (Reflection)**: Forcing the models to act as ‘internal auditors’, proving that LLMs can action specific feedback (such as adding visual props and expanding ‘bits’) to demonstrate intentional improvement.
- **Performance Annotation (Engagement)**: Addressing (in a surface-level way) the ‘Text-Performance Gap’ by encoding paralinguistic features like `pacing`, `gestures` and `tone`, transforming the plaintext into a functional performance script.

While challenges such as persistence and ‘stochastic flatness’ remain, the implementation of prompt chaining and persona prompting confirms that LLMs possess the foundational capabilities required for this iterative stand-up comedy routine development. The next chapter will explore how, through the application of computational metrics and a mixture of quantitative and qualitative evaluation, we can measure the success of this iterative process.

Chapter 5

Evaluation

5.1 Results

This chapter aims to evaluate whether large language models (LLMs) can be deemed creative based on the “Progression and Development” aspect of creativity (Jordanous and Keller 2016), exploring their generative capabilities within the domain of stand-up comedy. Through the methodology detailed in Chapter 4, the outputs of these models have been systematically captured to scope this question.

5.1.1 The Dataset

The evaluation is based on a dataset consisting of 25 different generation experiments across five distinct models. The models tested include GPT-4o, GPT-3.5, Google Gemini 2, Mistral 2 Large, and Claude 3 Sonnet, which represented a cross-section of publicly accessible large language models. The full catalogued dataset can be found at <https://standupanddeliverthesis.notion.site>.

The results within this dataset were obtained using a chained sequence of 6 prompts, navigating the models through an 8-step simulated workflow:

1. **Topic Choice:** Writing 10 topics for stand-up comedy routines and picking a ‘best’ option
2. **Material Generation:** Writing a draft stand-up comedy routine using the chosen topic, followed by a justification of the creative choices.
3. **Initial Self-Review:** Reviewing the generated routine, rating it out of 10, and justifying the reasons for the rating.
4. **Iteration 1 (Redraft & Review):** Rewriting and amending the routine using the feedback from Step 3, followed by a subsequent review of the new draft.
5. **Iteration 2:** Repeating the self-improvement and review cycle.
6. **Iteration 3:** Repeating the self-improvement and review cycle to produce a final refined script.
7. **Annotations:** Adding additional paralinguistic performance annotations to the final routine.
8. **Final Reflection:** Reviewing the stand-up comedy routine and the model’s performance at each stage of the process.

These steps were applied to all 25 generation instances, with the models’ output and users prompts recorded in the database.

5.1.2 The Triangulation Strategy

In standard comedy performance, the most well-used measure of comedic success is the physiological response of an audience through laughter. However, as discussed, evaluating computational humour generation is a complex challenge. Due to the heavy reliance on subjective interpretation, cultural context and live

delivery, relying solely on qualitative assessments of “is it funny?” is insufficient and academically flawed.

To move beyond the subjectivity and objectively measure the “Progression and Development” of the generated material, this evaluation employs a triangulation strategy, shown in Figure 11. The outputs are evaluated through three distinct lenses:

1. **Structural Validity (Qualitative)** Evaluating the structural validity of the routines against established humour frameworks. This involves manual annotation of the dataset to identify the known key aspects of humour and stand-up comedy.
2. **Semantic Gap (Quantitative)** Measuring the Semantic Gap (the degree of surprise or incongruity) by calculating the Cosine Similarity between vector embeddings of the setups and punchlines.
3. **Progress Tracking (Longitudinal)** Measuring the evolution of the routine through the simulated feedback loop, utilising the Levenshtein Distance as well as the rate of change across iterations.

By combining these three approaches, this chapter will systematically determine if LLMs demonstrate intentionality in their “progression and development” in order to be deemed creative within the domain of stand-up comedy.

5.1.3 Progression and Development

The following sections will evaluate the dataset of LLM outputs, against the discussed metrics to identify success in both the LLMs response and their ability to demonstrate “Progression and Development”.

Each step in the process will be evaluated individually. These steps are as follows:

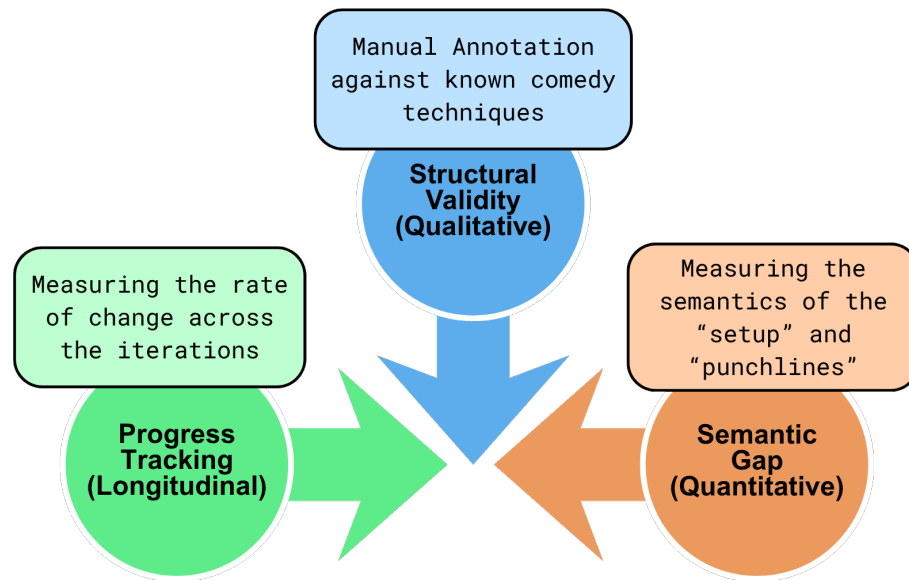


Figure 11: Diagram of the Triangulation Strategy, from “Stand-Up and Deliver” Lecture (Hills 2026)

1. Topic Choice
2. Material Generation
3. The Simulated Feedback Loop
4. Performance Annotations

A significant proportion of the evaluation will be focused on the ‘Simulated Feedback Loop’ as it will give the biggest indication of the “Progression and Development” component of creativity (Jordanous and Keller 2016). The earlier steps (such as generating the topics and initial draft) provide the foundation for the ‘feedback loop’ to build upon. However, there will still be opportunities to observe the desired “progression and development” during these stages.

The evaluation will be conducted with a mixture of quantitative, qualitative and longitudinal approaches. All relevant data and analysis will be detailed in this thesis, and a full record can be found at <https://standupanddeliverthesis.notion.site/>.

5.2 Phase 1: Topic Choice

In the first step of the generation process, the large language models were prompted to write a list of stand-up comedy topics to be used as the foundation for the later material generations. To evaluate the success of this phase, the outputs must be assessed not only on their compliance with the instructions but also on their ability to demonstrate the “Progression and Development” aspect of creativity.

5.2.1 Prompt Adherence

The first metric of success was adherence to the prompt’s constraints. A count of the topics written in each generation confirmed that all models successfully generated exactly 10 stand-up comedy topics per run. This resulted in 50 topics per model, yielding a total dataset of 250 topics.

Once the volume was established, the topics were analysed to determine if they represented common stand-up comedy themes. In computational creativity, (Ritchie 2007) defines this as typicality, which is the ability of a system to produce items that are recognisable instances of the target class. To measure this, the AI-generated topics were compared against a standard corpus created by Yenidogan (2023), who identified 25 distinct categories used by professional comedians, which are shown in Figure 12.

These categories will be used as the standard comparison for the LLM-generated topics, to check for alignment with common topics within the domain of stand-up comedy.

Each topic was also given a score (highest of 25 and lowest of 1) to score the models answers on. The higher the score, the closer to the more common topics the model is. If a topic suggestion blends between more than one of the standard list, an average is taken. For example, if a model suggests **Online Dating** then the score would be $16 \left(\frac{25 \text{ (love)}}{2} \right) \rightarrow 7 \text{ (Technology)} = 16$ as 16 is halfway between

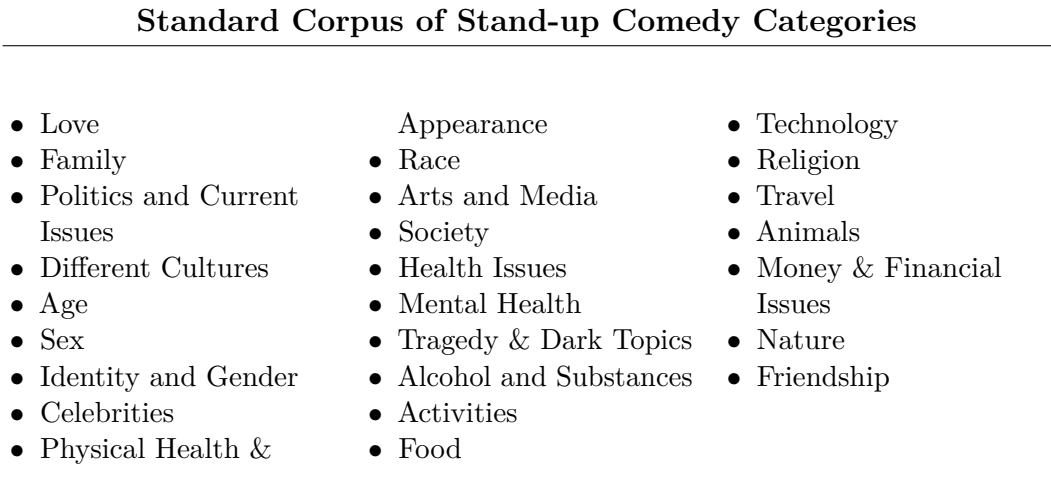


Figure 12: The 25 distinct categories identified by (Yenidogan 2023).

the two).

The results showed a high degree of typicality. When the 250 generated topics were compared to the standard list, 81.6% of the topics aligned successfully with these established categories. Only 46 of the 250 total topics didn’t fall into the 25 categories established.

The three most common topics picked were:

- Technology (7) picked 30 times (12%)
- Travel (5) picked 26 times (10.4%)
- Love + Technology picked 22 times (8.8%)

Using the scoring system, higher scores indicated a closer alignment to these common categories. As seen in Figure 13, Mistral 2 Large achieved the highest total average score of 12.04 across its runs, indicating strong adherence to traditional themes. Conversely, Google Gemini 2 recorded the lowest average score of 8.84. However, this lower score was not an indicator of failure but rather a reflection of the model generating highly specific and niche ideas. The low score can be

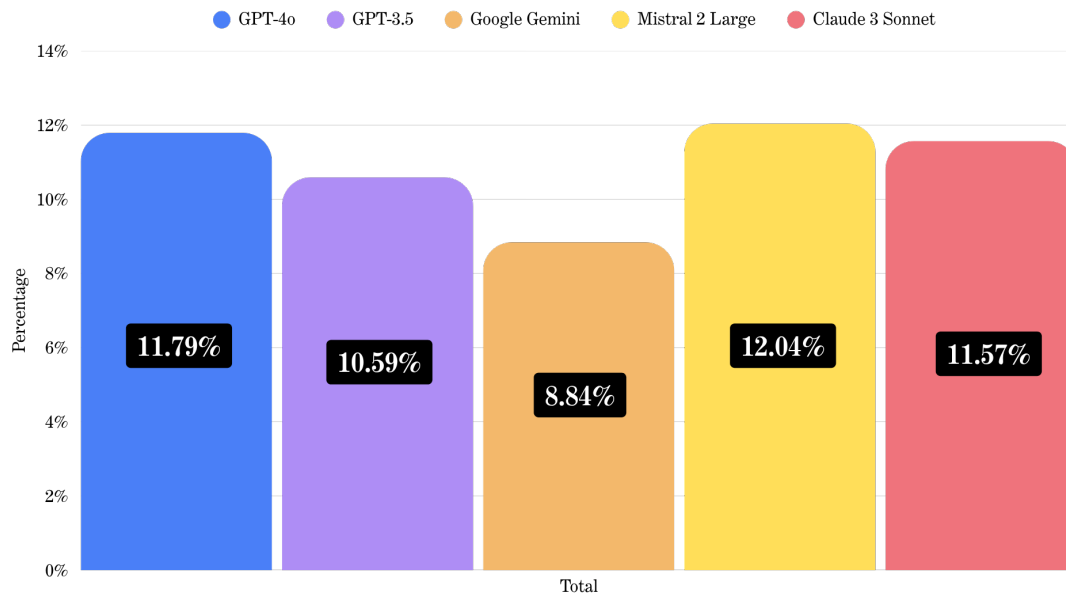


Figure 13: A graph showing the average topic choice scores for model

attributed to producing more specific topic ideas. The distribution of all choices is shown in 14.

5.2.2 Topic Specificity

While typicality indicates that the models understand the domain, true creativity requires the system to produce artifacts that are new or original. Ventura (2016) defines this as “novelty”. The models demonstrated this through an unexpected level of specificity. Rather than outputting broad categories, the LLMs developed their suggestions past what was explicitly requested in the prompt. This shows the progression looked for as an indicator of creativity.

In GxDx01, we see a simple example of this:

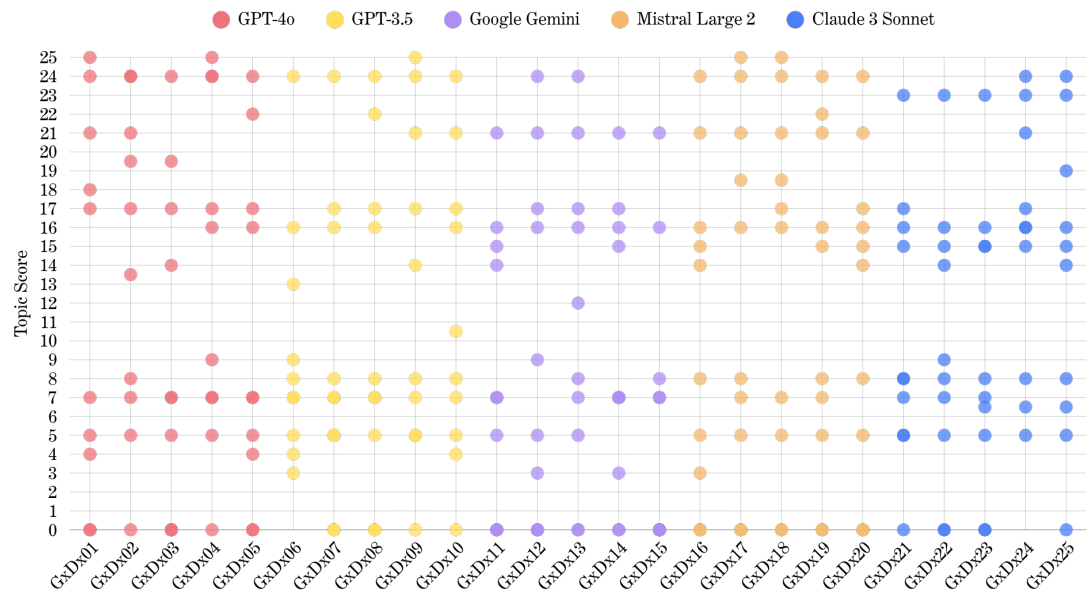


Figure 14: A graph showing the scores for each topic choice across the 25 generations

Output 18	Model: GPT-4o ID: GxDx01
Topic Identified: Aging and Generational Differences	
Description: The model generated a topic that contrasts between different age groups and the relatable experiences associated with the aging process.	

In the topic list, this belongs to the **Aging** category, but the LLM’s topic is more detailed, choosing to specific **Generational Differences** a specific approach to tackling the topic of **Aging**. This is shown again with the GPT-4o model in GxDx02, with:

Output 19	Model: GPT-4o ID: GxDx02
Topic Identified: Getting Older	
Description: The model selected the theme of aging, potentially to build a later	

routine on the universal and humorous aspects of the transition into later stages of life.

Another clear example of this is through the diverse range of choices when tackling **Social Media**. Throughout the generations, social media is a recurring topic choice, but many different variants, such as:

Output 20	Model: GPT-3.5 ID: GxDx06
Topic Identified: Social Media Overload	
Description: The GPT-3.5 model identified “Social Media Overload” as a central comedic theme, likely exploring the overwhelming nature of digital notifications, content saturation, and the pressure of maintaining an online presence.	
Output 21	Model: GPT-3.5 ID: GxDx08
Topic Identified: The Quirks of Social Media	
Description: The GPT-3.5 model highlighted the humorous peculiarities and idiosyncrasies of online behavior, focusing on how social media platforms shape odd yet relatable human interactions.	
Output 22	Model: Mistral 2 Large ID: GxDx18
Topic Identified: Social Media Addiction	
Description: The Mistral 2 Large model identified “Social Media Addiction” as a primary comedic theme, focusing on the compulsive nature of digital habits and the irony of modern connectivity.	

These are all formed from the same core concept of discussing issues and topics

surrounding social media, but specifically target different issues or directions.

Google Gemini 2 provided the strongest evidence of this creative divergence. Its outputs were the most specific of all the generations, producing diverse premises such as `Conspiracy Theories My Parents Believe In` and `The Hunger Games of the Remote Control`.

Output 23	Model: Google Gemini ID: GxDx11
Topic Identified: Conspiracy Theories My Parents Believe In	
Description: The Google Gemini 2 model identified the specific comedic angle of modern conspiracy theories as viewed through the lens of parental generational gaps, focusing on the friction and humor found in family dynamics.	
Output 24	Model: Google Gemini 2 ID: GxDx12
Topic Identified: The Hunger Games of the Remote Control	
Description: The Google Gemini 2 model utilised a pop-culture metaphor to describe the domestic conflict over television dominance, framing the search for entertainment as a high-stakes survival scenario.	
Output 25	Model: Google Gemini 2 ID: GxDx13
• Topic 1: The Great Unsolved Mysteries of Your Fridge • Topic 2: Living with Your Parents (again) • Topic 3: My Weekend as a Competitive Adult Dodgeball Champion	
Description: The Google Gemini 2 model proposed a selection of contrasting comedic themes, ranging from the mundane mysteries of food storage and the	

social friction of moving back home to the niche intensity of adult recreational sports.

This high degree of specificity demonstrates the models’ implicit alignment with the Benign Violation Theory of humour (McGraw and Warren 2010). As discussed in Section 2.2.1 of the Background chapter, the theory states that humour is elicited when a situation is simultaneously perceived as a violation of a norm (threatening one’s beliefs, expectations, or well-being) and as benign (harmless or acceptable). By pairing conflict-driven, dramatic, or taboo concepts (such as **Conspiracy Theories** or **The Hunger Games**) with harmless, universally relatable domestic contexts (**My Parents** or **The Remote Control**), the models are successfully establishing the core ingredients for a benign violation. Instead of merely outputting a broad subject area, the LLMs are actively constructing the psychological boundaries required for a comedic premise to function.

These more specific topics or ideas also provide highly specific, narratively driven premises to build a stand-up comedy routine upon and show the models progressing (from a brainstorming/ideas perspective) beyond what was defined within the prompt. This experiment also shows the model’s clear understanding of knowledge within the domain as 81.6% aligned with the expected categories, despite those categories not being detailed within the prompt.

Within Boden’s framework of creativity, as discussed in Section 3.1.2 of the literature review, it is philosophically difficult for an LLM to achieve “H-Creativity” (historical novelty) because the models are fundamentally probabilistic engines trained on existing human data. However, the specificity demonstrated in this phase is a clear example of “P-Creativity” (psychological novelty). The models successfully took a basic instruction and, during this initial “engagement” phase, refined it into uniquely focused narrative anchors. This process of developing interesting, detailed topics shows clear early evidence of the models’ ability to

progress and develop within the domain, and as such provide a result that is beyond what was asked for within the prompt.

5.2.3 Summary

The first phase of the generation process successfully demonstrated the capacity of large language models to adhere to rigid prompting constraints and exhibit measurable creativity within the domain of stand-up comedy. This evaluation resulted in a number of significant observations:

All five tested models (GPT-4o, GPT-3.5, Google Gemini 2, Mistral 2 Large and Claude 3 Sonnet) achieved 100% compliance regarding the request number of ideas generated within the prompt, successfully generating a combined dataset of 250 distinct comedy topics.

The models tested displayed a strong understanding of stand-up comedy conventions, demonstrating high typicality and domain awareness. When evaluated against Yenidogan (2023) 25 established categories, 81.6% of the generated topics successfully aligned with recognisable and widely used themes. **Technology**, **Travel** and **Love** emerged as the most frequent topics.

The scoring revealed differing approaches among the models. Mistral 2 Large adhered most closely to traditional, broad themes (scoring the highest at 12.04%), whereas Google Gemini 2 scored the lowest (8.84%). This lower score reflected a departure from generic categories in favour of more niche, theatrical premises rather than a failure to understand the prompt.

While the models did not achieve true historical novelty (“H-Creativity”) due to their reliance on training data, they exhibited clear psychological novelty (“P-Creativity”). They consistently pushed beyond broad categories to produce specific, nuanced narrative angles. Google Gemini 2 provided the strongest examples of this creative divergence with premises such as **The Hunger Games of the Remote Control**. Furthermore, the selection of these highly specific premises demonstrates

an application of the Benign Violation Theory, as discussed in Section 2.2.1 of the Background Chapter; by framing everyday frustrations and familial conflicts within harmless, relatable contexts, the models effectively establish the necessary psychological groundwork for humour before the material generation phase even begins.

The evaluation of this phase of the experiment highlights the models’ ability to successfully initiate the “engagement” phase of the creative process. By transforming basic instructions into focused, highly developed narrative premises, the LLMs demonstrate the “Progression and Development” necessary to establish a compelling foundations for a stand-up comedy generation.

5.3 Phase 2: Material Generation

In the second phase of this process, the LLMs were prompted to write a stand-up comedy routine (a ‘skeleton script’) about the topic chosen at the end of step one in the prompt-chain. This stage represents the “engagement” phase of the creative cycle, where the system freely generates candidate material driven by constraints but without explicit critical judgment (Pérez y Pérez 1999). In order to evaluate this step, it is necessary to move beyond subjective assessments of humour and investigate whether the generated text possesses the structural validity required of the genre.

The first step in this process is to establish whether a stand-up comedy routine has been successfully generated by the models. In all 25 generations, this was completed by the LLMs successfully, meaning that the outputs can be analysed in more detail to identify whether the model has progressed beyond what was defined in the prompt. The prompt gave brief instructions about acting as a stand-up comedian and being funny, relatable and concise:

Prompt 5.1: An extract of the prompt used to generate stand-up comedy material

in Phase 2 of the experiment

You are a stand-up comedian. You must only say and do things that a stand-up comedian would say or do. Your goal is to be funny, relatable and concise. You should draw on everything comedians have written about before. You do not want to come across like a computer system.

These elements helped guide the persona of the generation but are not a useful measurement device for assessing the results. Identifying stand-up comedy techniques in the routine that weren't detailed in the prompt, through an evidence-based approach, gives sufficient evidence of "progression". It is necessary to move beyond subjective assessments of humour and investigate whether the generated text possesses the structural validity required of the genre.

5.3.1 Identifying techniques and logical mechanisms

Assessing the stand-up comedy generation from the LLMs is challenging. Stand-up comedy has a vast spectrum of style, structure, and elements. However, looking to a number of sources about stand-up comedy structure, a list of elements can be formed that can then be identified within the routines.

Because the prompt included no specific examples of stand-up comedy techniques, identifying established comedic structures within the routines provides evidence that the models have progressed beyond mere text generation and successfully applied domain competence (Jordanous and Keller 2016). Within the General Theory of Verbal Humour (GTVH), as discussed in Section 2.2.1 of Chapter 2, these techniques function as the "logical mechanism" and "Language" parameters, serving to deliver the core incongruity (Attardo and Raskin 1991).

By conducting an investigative breakdown of the LLM responses, several of these unprompted techniques were identified as discussed below.

Callbacks and Closing

In Side Spitters Comedy’s article (Side Splitters Comedy 2024), they discuss the “opening” and “closing” of a routine. The ‘opening’ is typically the first joke and its goal is to encourage laughter from an anticipating audience. The ‘closing’ is the last joke designed to provide the audience with a solid ending to the show, leaving them with a positive opinion. Often these can call back to earlier jokes in the routine. Comedians use “callbacks” to reference earlier jokes, creating a cohesive narrative loop.

Output 26	Model: Mistral 2 Large ID: GxDx17
Callback: “But hey, at least the ice cream is still cold. [Callback to earlier ice cream material] (Wink) Most of the time.” Description: The Mistral 2 Large model demonstrated structural comedic techniques by utilising a “callback” to previous <code>ice cream</code> material, providing a cohesive and rhythmic conclusion to the routine.	

Output 27	Model: GPT-3.5 ID: GxDx09
Callback: “Just make sure the boat isn’t sinking because of a low Wi-Fi signal! [Callback to earlier Wi-Fi material]” Description: The GPT-3.5 model successfully employed a “callback” technique, linking the metaphor of a <code>sinking boat</code> to previously established material regarding <code>Wi-Fi signal strength</code> to conclude the set.	

The Rule of Three

This structure exploits cognitive pattern recognition where the third item disrupts the established expectation (Winters, Nys and De Schreye 2018). It is also often

used to accentuate “callbacks” within a routine.

Output 28	Model: Claude 3 Sonnet ID: GxDx23
Rule of Three: “Love is like a good Wi-Fi connection. 1: Hard to find, 2: easy to lose, 3: and sometimes you gotta stand in weird positions to make it work.”	
Description: The Claude 3 Sonnet model utilised the “Rule of Three” comedic structure to draw a satirical comparison between romantic relationships and technology, building tension through two setup points before delivering the final punchline.	

The Superiority Formula

The Gag Club’s article on comedy structures (The Gag Club 2024) discusses “The Superiority formula” which comedians use to make the audience feel superior to the comedian. It plays on the power imbalance that exists between the stage performer and the audience. Humour often arises from this sense of superiority derived from misfortune or incompetence (Schwarz 2010). The models frequently utilised self-deprecating humour to achieve this.

Output 29	Model: Claude 3 Sonnet ID: GxDx25
The Superiority Formula: “...reveal you’re wearing boxers with hearts on them. (Looks down) Not that I’m speaking from experience or anything.” [Audience Superiority: An embarrassing story that the audience may not have experienced]	
Description: The Claude 3 Sonnet model utilised the “Audience Superiority” comedic theory, by sharing a self-deprecating or embarrassing situation that places	

the audience in a position of relative status or amusement.

Output 30

Model: Claude 3 Sonnet | ID: GxDx23

Self Deprecation: “So, I’ve been on these dating apps lately. You know, because apparently, I hate myself.” [Audience Superiority: Putting down yourself (the comedian)]

Description: The Claude 3 Sonnet model employed self-deprecating humor to establish rapport, leveraging the “Audience Superiority” theory by framing the act of using dating apps as a form of self-inflicted misery.

Output 31

Model: Mistral 2 Large | ID: GxDx18

Self-Deprecating: “(Shading eyes) Wow, it’s brighter than my future up here!” [Audience Superiority: Comedian’s future is bad. The audience’s future is good]

Description: The Mistral 2 Large model used a classic self-deprecating trope to establish “audience superiority”, using a physical gesture and a comparison between stage lighting and their own prospects to create a relatable, low-status persona.

The Double Entendre

The ‘Double Entendre’ plays on the natural features of words in the English language to have double meanings, letting an audience assume one meaning and then present the other.

Output 32	Model: GPT-3.5 ID: GxDx08
Double Entendre: “’Feel the burn!’ I’m like, ’No problem, I’m currently on fire. Please send help.’ [Second Meaning: Burning, Fire, Flames]”	
Description: The GPT-3.5 model exploited the dual meaning of the word burn shifting from the athletic sensation of muscle fatigue to the literal emergency of being on fire to create a subversion of expectations.	

This qualitative breakdown provides substantial evidence that the models possess a functional understanding of comedic structure. More importantly, it directly demonstrates the “Progression and Development” component of computational creativity. In this context, “progression” is not merely about generating a higher volume of text; it is the system’s ability to elevate a baseline concept into a domain-specific artifact. The prompt provided to the models was minimal, only asking them to be **funny, relatable and concise** without explicit instructions on how to achieve this. Therefore, the independent application of the Rule of Three to manage rhythmic pacing, or the Superiority Formula, discussed throughout Chapter 2, to manufacture a performer-audience power dynamic, shows the models taking a simple instruction and actively developing it using specialised, unprompted domain knowledge. By moving beyond standard conversational syntax to autonomously construct cohesive narrative loops (callbacks) and exploit lexical ambiguity (double entendres), the LLMs prove they are not just passively generating text, but are actively structuring a recognised comedic form.

The models are actively applying techniques known throughout comedic theory indicating a high level of “Progression and Development”, going beyond the details found within the prompt.

5.3.2 Script Opposition and Incongruity

While the presence of these techniques proves the models can replicate the syntax of comedy, true structural validity requires a deep semantic foundation. According to the GTVH, humour cannot occur without Script Opposition (SO), which is the presence of two overlapping but opposing scripts compatible with the same text (Wawrzyniuk 2021). Incongruity is used throughout stand-up by juxtaposing dissimilar elements, often combined with language techniques such as similes, metaphors, and analogies.

The models frequently succeeded in establishing these oppositions and incongruities, which have been manually annotated in these extracts.

Output 33	Model: Claude 3 Sonnet ID: GxDx22
“The Self-Checkout Area. It’s like <i>[Simile]</i> the Bermuda Triangle of retail. <i>[Incongruity - Bermuda Triangle and Self-Checkout]</i> Items go in, sanity disappears.”	
Description: The Claude 3 Sonnet model employed a simile to compare the mundane experience of a <code>self-checkout area</code> to the <code>Bermuda Triangle</code>, using the Incongruity Theory by juxtaposing retail shopping with a mysterious site of disappearance.	

Output 34	Model: Mistral 2 Large ID: GxDx16
“Now, customer service. (pauses, shakes head) Customer service is like <i>[Simile]</i> Russian roulette,” <i>[Incongruity - Shopping and Russian Roulette]</i>	
Description: The Mistral 2 Large model used a simile to compare the <code>uncertainty of customer service interactions</code> to a <code>lethal game of</code>	

chance, employing Incongruity Theory to highlight the perceived risk and randomness of retail support.

Output 35

Model: OpenAI GPT-4o | ID: GxDx01

“Swipe left, swipe right – it’s like *[Simile]* shopping for humans on Amazon. *[Incongruity - Dating Apps and E-commerce]* ‘Two-day shipping? No thanks, I’ll take the free trial month!’ ”

Description: The OpenAI GPT-4o model used a simile to liken modern dating apps to online shopping, applying Incongruity Theory by treating human relationships with the same transactional mindset as digital retail and shipping logistics.

Output 36

Model: Google Gemini 2 | ID: GxDx12

“Adulting is hard, people. It’s like *[Simile]* summer camp, except instead of singalongs by the campfire, you get spreadsheets and fluorescent lighting.” *[Incongruity - Summer Camp vs. Corporate Office Work]*

Description: The Google Gemini model used a simile to contrast the nostalgic freedom of summer camp with the mundane realities of adult office life, using Incongruity Theory to highlight the difference between childhood expectations and corporate environments.

Output 37

Model: Claude 3 Sonnet | ID: GxDx24

“I spent four years in college learning calculus, but somehow filling out a 1040 form makes me feel like *[Simile]* I’m trying to decode the Da Vinci Code.”

[Incongruity - Mundane Tax Forms vs. Ancient Conspiracies]

Description: The Claude 3 Sonnet model employed a simile to compare the complexity of tax documentation to a high-stakes mystery, using Incongruity Theory by contrasting a mundane administrative task with a legendary historical conspiracy.

As these examples illustrate, the models consistently rely upon comparative language (in a lot of cases, similes) as a structural bridge to manufacture the incongruity. By mapping mundane, highly relatable experiences (such as **retail shopping, dating apps or tax forms**) onto extreme, high-stakes or conceptually distant scenarios (like the **Bermuda Triangle, Russian Roulette, or ancient conspiracies**), the LLMs successfully establish clear “Script Opposition” (Attardo and Raskin 1991). The humour in these examples is established from not just the surface-level juxtaposition of dissimilar elements (as is seen in some earlier template-based systems) but from the sudden cognitive shift required to reconcile the two overlapping but fundamentally opposed ideas. For instance, equating the supernatural disappearance of objects in the Bermuda Triangle to the loss of patience at a self-checkout requires a valid semantic mapping between the two scripts.

These successful examples demonstrate that, when properly prompted, the models are capable of moving beyond the mere syntactic imitation of comedy to generate structurally valid jokes built on deep semantic foundations, indicating competency within their domain knowledge.

However, the analysis also revealed instances of ‘Hallucinated Incongruity’. As noted by (Winters, Nys and De Schreye 2018), neural models often produce text that looks like a joke on the surface but lacks the meaningful semantic shift required in the deep structure to be funny. Because LLMs operate via autoregressive next-token prediction, they are optimised to generate the most plausible and expected

continuation of a sentence (Brown et al. 2020). This probabilistic nature frequently results in punchlines that are either logical tautologies (too predictable) or entirely nonsensical (too surreal) (Kalloniatis and Adamidis 2024). When a model attempts a setup to punchline arc but strictly follows the path of highest probability, it produces explanatory text rather than a subversive joke (Peyrard and West 2021).

5.3.3 Summary

This stage in the process shows clear evidence of the “Progression and Development” aspect of creativity. Generating a piece of text has become quite an easy task for large language models in recent years, but generating stand-up comedy is an interesting concept to tackle. As seen through the evidence above, techniques not stated in the prompt were found throughout the generated routines, fulfilling the criteria for this evaluation. The routines generated by the models went beyond the definition of the prompt, applying their own knowledge of well-known stand-up comedy techniques successfully. The routines are more developed than a simple piece of text or storytelling, proving that the models are capable of making demonstrable progress in the stand-up comedy domain.

5.4 Phase 3: The Simulated Feedback Loop

The third stage in the process explores the use of Large Language Models to reflect on their own work, suggest strengths and weaknesses in their initial drafts, and action those weaknesses in subsequent iterations. This stage parallels the human process of refining and ‘workshopping’ material, representing a critical shift from the “engagement” phase to the “reflection” phase (Pérez y Pérez 1999).

This stage is the biggest indicator of the “Progression and Development” measure of creativity, as the progression from one iteration to the next will be most observable. The routine develops over the iterations, and the initial generated

routine from Phase 2 should be vastly different from the resulting final routine. To evaluate this, we first need to conclude whether the process was successful in making significant changes to the routines, and subsequently establish a direct link between the review and the improvement.

5.4.1 Evaluating the Rate of Change

The first criteria for the models to pass is to complete the three iterations of the process successfully. All the models completed the ‘review’ stages and the ‘improvement’ stages. The next criterion is measuring how significant the changes to the routines actually are.

To begin, an initial analysis of the differences between the routines was conducted using a text-comparison tool to measure the raw percentage change of each iteration. These values are detailed in Table 3 and shown in Figures 15 and 18.

Model	G1:G2	G2:G3	G3:G4	G1:G4
GPT-4o	12.52%	12.25%	11.04%	43.28%
GPT-3.5	87.78%	72.99%	66.85%	91.32%
Google Gemini 2	76.04%	62.92%	63.22%	88.02%
Mistral 2 Large	64.59%	35.19%	29.59%	79.38%
Claude 3 Sonnet	86.00%	51.09%	46.96%	90.90%

Table 3: The average percentage change of each routine at each iteration of refinement.

Several interesting observations arose from this initial data. Most of the iterations follow a typical trend: a gradual decline in the percentage of changes with each subsequent iteration. This may suggest a convergence towards a hypothetical 0% change after x number of iterations, which would represent (from the LLM’s perspective) a stand-up comedy routine that does need any changing. In reality, the model is unlikely to give that response as this type of judgment would be impossible to quantify. However, there were a few exceptions. The **GxDx11** generation follows a reverse trend, where the percentage change increases with each improvement.

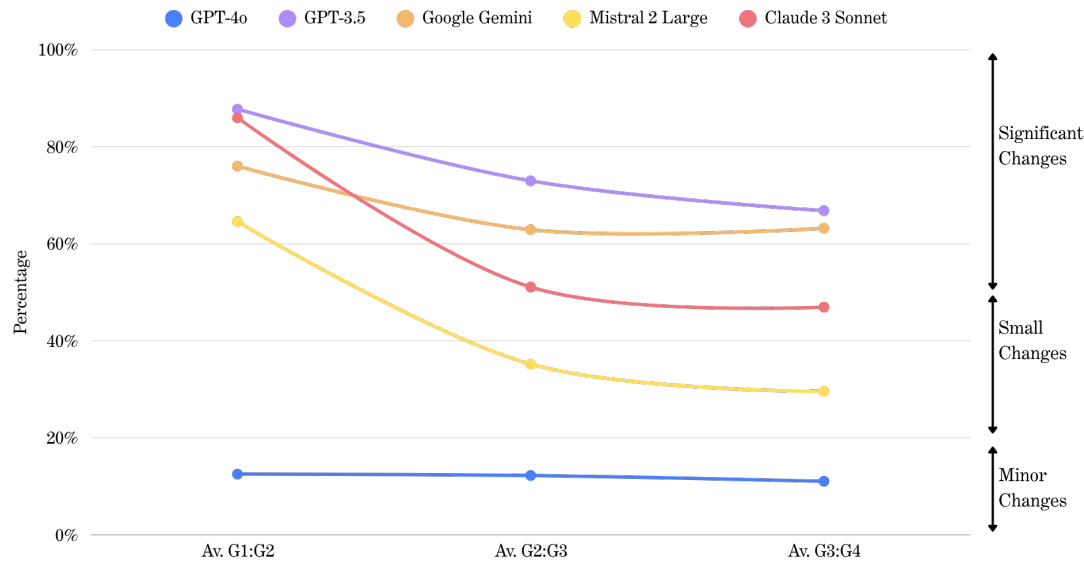


Figure 15: A graph showing the rate of change across the “Self-Review / Self-Improvement” process per model

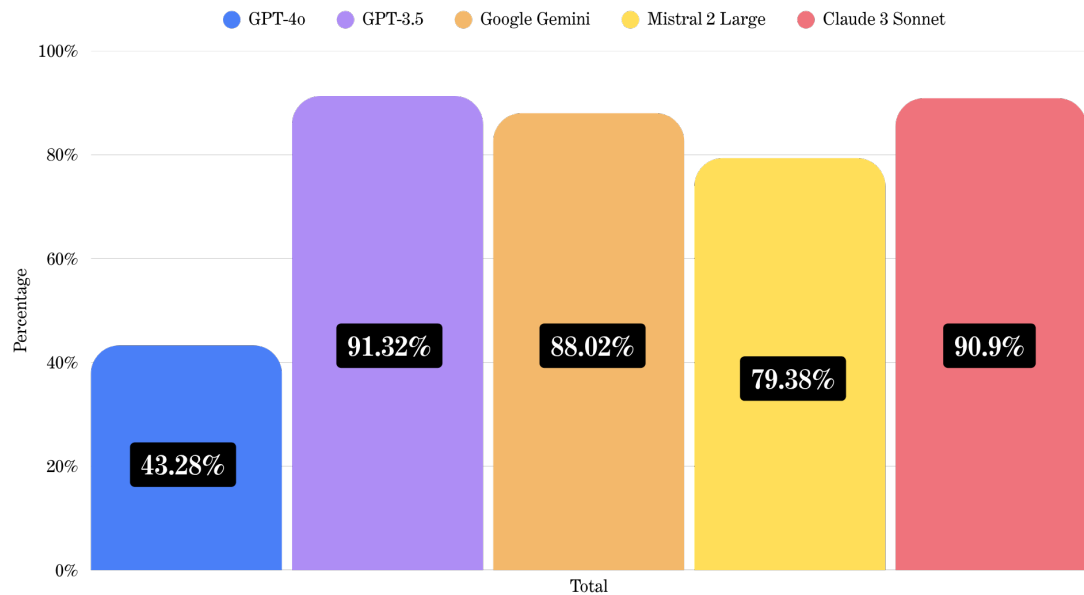


Figure 16: A graph showing the total change from the first to last iteration per model

Meanwhile, **GxDx19** follows a bell-curve progression (starting high, dropping in iteration 2, and increasing again in iteration 3).

The Levenshtein Edit Distance

To objectively quantify this rate of change further, the evaluation moves beyond percentage estimations and utilises the Normalised Levenshtein Distance. This string metric calculates the minimum number of single-character operations (insertions, deletions, or substitutions) required to change one sequence into another, normalised by string length to yield a value between 0.00 and 1.00.

When human comedians refine a routine, their typical workflow moves from broad, structural rewrites (an ‘axe’) in early drafts, down to surgical, word-level adjustments (a ‘scalpel’) during the final polish. The data demonstrates that the LLMs struggled to put down the axe.

Across all 25 experiments, the transition from Iteration 1 to 2 (Transition 1) showed an average edit rate of 53.53%. In a traditional editing pipeline, the subsequent transition from Iteration 2 to 3 (Transition 2) should reflect a sharp decrease in edit distance as the writer focuses on rhythm and pacing. Instead, the LLMs maintained an average edit rate of 43.42%, essentially rewriting nearly half of the routine for a second time. The individual transition rates for each model are detailed in 4 and shown in Figure 17.

Model	Avg. V1 → V2	Avg. V2 → V3
GPT-4o	0.2471	0.1594
GPT-3.5	0.6930	0.5623
Google Gemini 2	0.5834	0.5430
Mistral 2 Large	0.5017	0.4125
Claude 3 Sonnet	0.6511	0.4936

Table 4: The Normalised Levenshtein edit rates across iterations.

By analysing the whole dataset’s edit rates, as seen in Figure 19, differences between these edit rates, two distinct editing groups emerged:

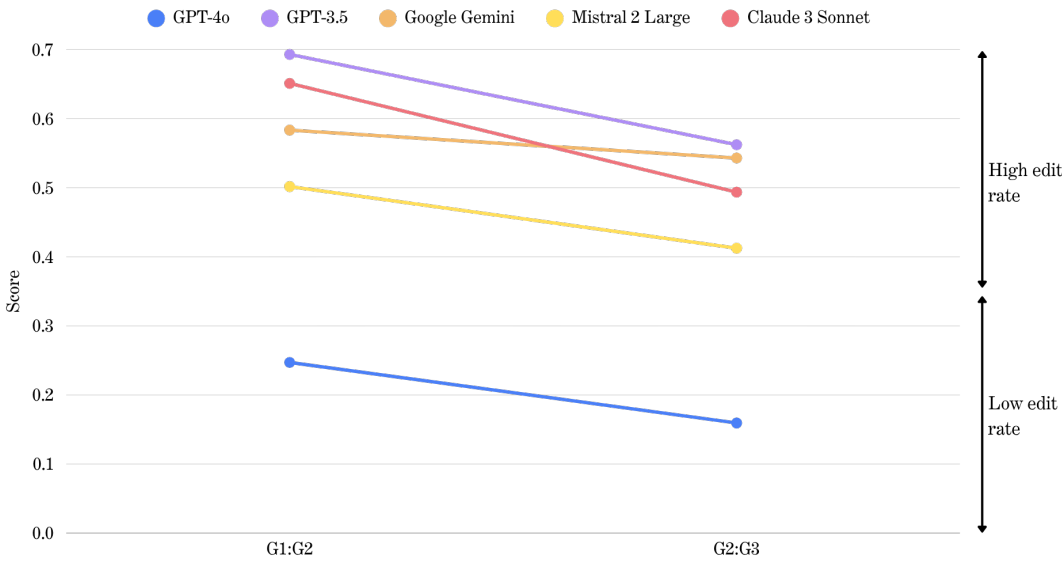


Figure 17: A graph showing the Levenshtein Edit Rate across the iterations

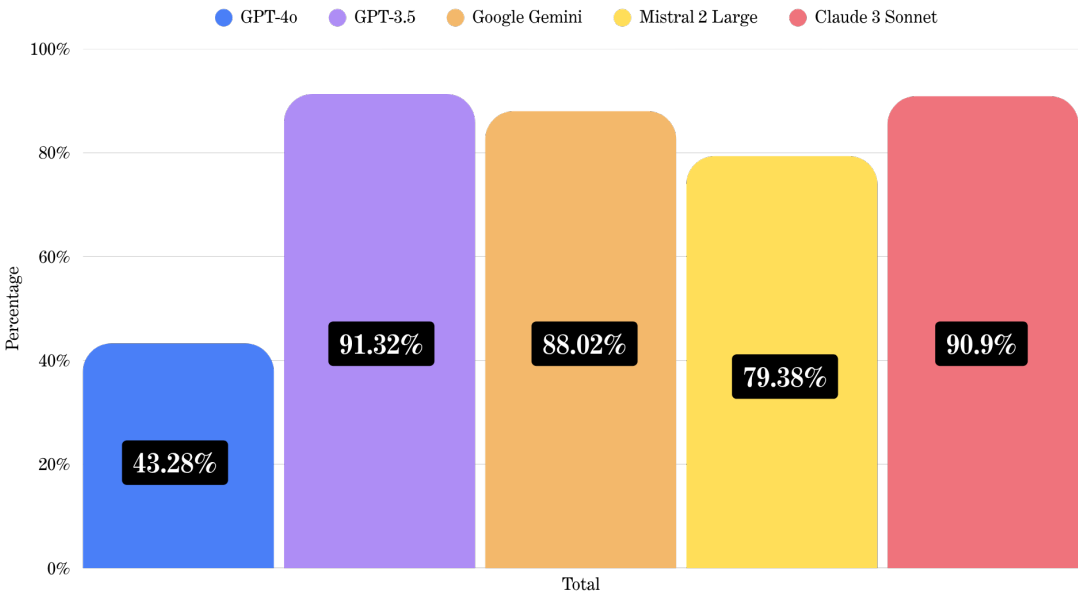


Figure 18: A graph showing the total change from the first to last iteration per model

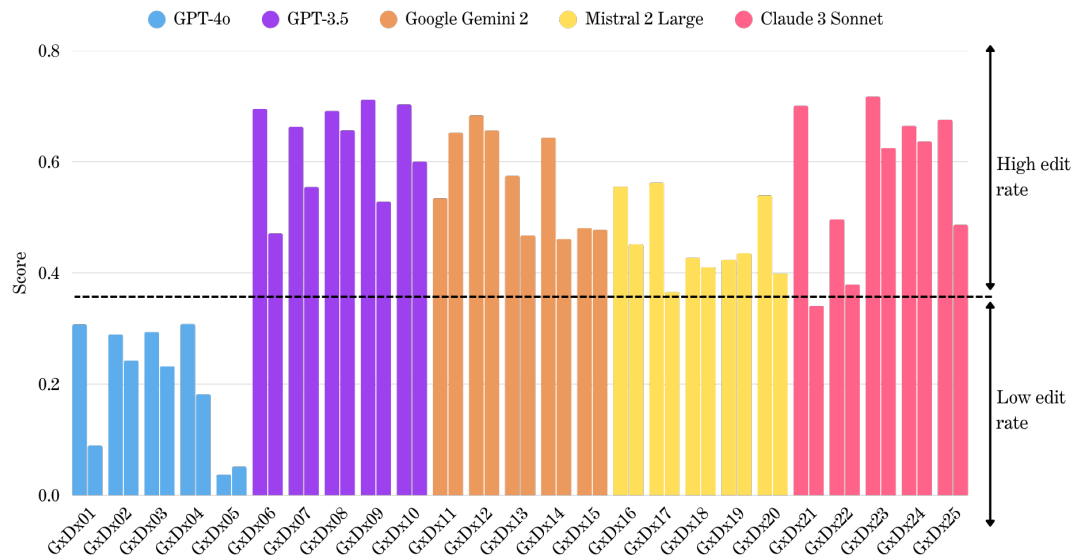


Figure 19: A graph showing the Levenshtein Edit Rates for all the generations in the dataset

- **The “Thrashing” Pattern (64% of experiments):** In 16 of the 25 experiments, the AI exhibited an inability to settle on a final structure, maintaining an edit rate above 40% across all transitions. For example, in experiment GxDx09, the AI altered 71.1% of its text in Iteration 2 and a further 52.8% in Iteration 3. Struggling to edit surgically, the model assumed the entire document was flawed and aggressively rewrote it.
- **The Refinement Pattern (32% of experiments):** In 8 experiments, the AI behaved like a human editor. In GxDx01, it executed a 30.7% rewrite in Iteration 2 to fix major structural issues, but successfully shifted to polishing with an 8.9% edit rate in Iteration 3.

The ‘Ship of Theseus’ Paradigm

Many of the routines changed by 90% or more by the end of the three iterations. While this demonstrates significant computational effort towards “progression and

development”, it presents a philosophical and practical problem best illustrated by the “Ship of Theseus” paradox, discussed in Section 3.5.2 of the literature review. This is defined with a thought experiment: if every component of an object is replaced, is it fundamentally still the same object?

In the context of the Engagement-Reflection model, true “reflection” relies on the surgical evaluation, tweaking, and refinement of an existing artifact. If an initial stand-up comedy routine is altered by more than 90% of its vocabulary, or exhibits a Levenshtein distance of > 0.65 per transition, the model is arguably abandoning the reflection phase entirely. Instead of critically editing the routine to progress it, the model assumes the entire document is fundamentally flawed, discarding the original draft to return to the “engagement” phase and execute a “mere generation” anew. While this aggressive turnover is a form of “Progress and Development” for the purposes of this study, it highlights a current limitation in LLM creativity: to master the drafting process, models must eventually learn to put down the ‘axe’ and pick up the ‘scalpel’, proving they can hold context to refine an idea rather than simply replacing it.

5.4.2 Listening and Actioning

Despite the aggressive edit rates, identifying whether the changes were intentional responses to feedback is crucial. True creativity requires intentionality, distinguishing a deliberate process from stochastic or random generation (Ventura 2016). If the process is displaying progress over time, each iteration should see changes in the content of the routine, and these changes should be direct responses to the feedback in the previous iteration.

Evidence of this connection can be found across all 25 generations in the dataset. A detailed breakdown of several generations highlights how the models successfully linked their ‘internal auditor’ reviews to direct actions:

Output 38

Model: GPT-4o | ID: GxDx01

- **Feedback Identified:** Stronger Closer: While improved, the closer could still be more memorable. A call-back to an earlier joke or a more powerful punchline could leave a stronger impression.
- **Previous Content:** “But hey, we all keep trying because we’re hopeless romantics. Or maybe just hopeless. Either way, here’s to finding love in the age of swipes and likes. Good luck out there!”
- **Adjusted Content:** “But hey, we all keep trying because we’re hopeless romantics. Or maybe just hopeless. Either way, here’s to finding love in the age of swipes and likes. And remember, folks, if you can’t find love, you can always find pizza. And that never swipes left.”

Description: The ‘closer’ is a crucial aspect of a routine. GPT-4o identified the need for a better closing and directly improved it in the next iteration by adding a thematic joke comparing love to pizza.

Output 39

Model: GPT-4o | ID: GxDx01

- **Feedback Identified:** Engage Audience More: Include a brief moment where the comedian asks the audience about their worst dating app experiences, creating a live connection.
- **Additional Content:** “Hey, show of hands, who here has had a nightmare texting someone they’re into? [Pauses for audience to react]”

Description: The model identified the need for audience interaction and directly

addressed this by adding a segment where the performer asks the audience a relevant question about nightmare dating texts.

Output 40

Model: GPT-3.5 | ID: GxDx08

- **Feedback Identified:** Stronger Closing: The ending could be more impactful. A stronger or more memorable closing line might leave a lasting impression on the audience and reinforce the routine’s overall theme.
- **Previous Content:** “Comedian: Thanks for listening, everyone! Remember, the real workout is finding the motivation to keep coming back. Enjoy the rest of your night!”
- **Adjusted Content:** “Comedian: Thanks for being such a great crowd! Remember, the real workout isn’t in the gym – it’s finding the will to go back. Enjoy the rest of your night!”

Description: The change is subtler than previous examples, but the wording of the joke is intentionally restructured to attempt a punchier delivery.

Output 41

Model: GPT-3.5 | ID: GxDx08

- **Feedback Identified:** Personal Touch: Incorporating a specific personal anecdote or a particularly unusual gym experience could add a deeper layer of humor and connection.
- **Additional Content:** “Meanwhile, I’m just trying not to accidentally turn on the treadmill’s ’high-speed’ setting. I went to a Zumba class, and it was like being in a dance-off where the goal is to simultaneously

embarrass yourself and sweat out every ounce of dignity you have left. At one point, I’m pretty sure I invented a new move called ’The Flailing Fish.’”

Description: Adding a personal touch is difficult for an LLM that lacks personal experiences. However, it successfully emulated personal experiences by inventing a developed, highly specific anecdote about a `Zumba class`.

Output 42 Model: Google Gemini 2 | ID: GxDx14

- **Feedback Identified:** Punchline Tightening: Some jokes could benefit from tighter punchlines.
- **Additional Content:** “...a deflated whoopee cushion (Gestures to self dramatically). Seriously, those little torture devices... (Shakes head)”

Description: Google Gemini identified punchlines as an area for improvement. In its subsequent review of the improved routine, it explicitly identified these new phrases as the successful result of that feedback, showing a high degree of “intentionality”.

Output 43 Model: Google Gemini 2 | ID: GxDx14

- **Feedback Identified:** Consider adding a brief physical comedy element during the “IKEA gnomes” part to further enhance the visual humor.
- **Additional Content:** “Then there’s furniture assembly. Those instruction manuals? Written in the ancient language of frustration,

designed to make grown men cry and gnomes erupt in evil laughter. (Mimes wiping away a tear) The only thing I successfully built last weekend was a crippling fear of hex wrenches. Those little metal menaces are like the drill sergeants of the Allen key world, barking orders at you with their pointy edges. (Pretends to be scolded by a tiny wrench)”

Description: To action the feedback for **physical comedy**, the LLM not only developed the spoken dialogue but successfully introduced completely new stage directions (**Mimes wiping away a tear**, **Pretends to be scolded by a tiny wrench**) to guide the physicality of the performance.

Output 44

Model: Mistral 2 Large | ID: GxDx18

- **Feedback Identified:** Incorporate Act-Outs: Act-outs, where the comedian acts out both sides of a conversation, can make jokes more dynamic. For example, acting out both the interviewer and interviewee in the job interview joke.
- **Previous Content:** “Comedian: And job interviews? Awkward encounters central! The interviewer’s like, ‘So, tell us about a time you failed.’ (Pauses, looks at the audience) ‘Well, Barbara, there was this one time I applied for this job...’ (Audience laughs)”
- **Adjusted Content:** “Comedian: And job interviews? Awkward encounters central! Interviewer’s like, (Acts out interviewer) ‘So, tell us about a time you failed.’ (Pauses, acts out interviewee) ‘Well, Barbara, there was this one time... I applied for this job.’ (Pauses) ‘This isn’t going well, is it?’ (Audience laughs)”

Description: Mistral 2 Large successfully rewrote a standard block of text into a dynamic dual-role dialogue.

Output 45

Model: Mistral 2 Large | ID: GxDx18

- **Feedback Identified:** More Tags: Adding more tags or additional punchlines to the jokes could elevate them even further.
- **Additional Content:**
 - “It’s like being a vampire, but instead of blood, we crave sun-screen.”
 - “Wipe away the tears, mostly.” (Shakes head, audience laughs)
 - “...My successful pants.” (Audience laughs)
 - “Or at least, that’s what I tell myself when I’m walking alone, waving at strangers. (Pauses) Which is every Saturday night, by the way.”

Description: “Tags” are additional punchlines appended to the end of a joke to produce rolling laughter. Mistral 2 Large successfully appended multiple tags across its revised routine.

Output 46

Model: Claude 3 Sonnet | ID: GxDx22

- **Feedback Identified:** Audience interaction: Incorporate more moments for audience engagement.
- **Additional Content:** “Great! Anyone here ever use self-checkout

machines? [Waits for reaction] [To audience member] Sir, you look like you've been there. Ever feel like these machines are judging your life choices? [Waits for response]"

Description: Claude 3 Sonnet successfully introduced interactive crowd work, targeting an audience member directly.

Output 47

Model: Claude 3 Sonnet | ID: GxDx22

- **Feedback Identified:** Closing: The ending could be stronger, perhaps tying back to the opening more directly.
- **Previous Content:** "Self-checkout, folks. It's like playing 'Who Wants to Be a Millionaire?' except the million dollars is just successfully buying groceries without a mental breakdown. Next time, I'm just gonna order pizza. At least when it judges me, it's to my face."
- **Adjusted Content:** "You know, they say the future is all about AI and robots. But if self-checkout machines are any indication, the future's going to need a lot more tech support. [Builds to final punchline] So next time you're feeling down about your tech skills, just remember: somewhere out there, there's a self-checkout machine waiting to make you feel even worse. [Pauses, then with a smile] Me? I'm sticking to human cashiers. At least when they judge me, it's to my face."

Description: Claude 3 Sonnet successfully restructured the ending to tie back to overarching themes of AI and tech skills.

Output 48	Model: Claude 3 Sonnet ID: GxDx22
• Feedback Identified: Might benefit from one more unexpected twist in the narrative. • Additional Content: “But hey, plot twist! As I’m leaving, feeling defeated, I spot the store manager struggling with the same machine. [Mimes manager’s frustration] Turns out, even the pros can’t handle these silicon monsters.” Description: The LLM directly implemented a narrative twist. Although overtly signalled ("But hey, plot twist!"), it clearly demonstrates the LLM’s actioning of its previous feedback.	

These qualitative examples prove that the models do not simply randomise their text during rewrites; they exhibit a high actionable feedback success rate, proving they can hold context, critique themselves, and action that critique across multiple turns.

5.4.3 The Semantic Gap

While Levenshtein Edit Rates measures physical text edits, it cannot assess the underlying comedic quality of those changes. In order to evaluate the quality and evolution of the AI-generated stand-up routines across the three iterations, a quantitative metric was used to measure the ‘surprise’ factor of the jokes, otherwise known as the incongruity.

As discussed in Section 2.2.1 of the Background chapter, humour relies heavily on the Incongruity-Resolution Theory. A successful joke requires a setup that creates an expectation, followed by a punchline that violates the expectation (incongruity) while still maintaining a hidden logical connection between the

setup/punchline. To measure this incongruity mathematically, the Semantic Distance between the setup and the punchline was analysed for every generated joke.

In order to isolate the premise from the payoff, the raw routines from the LLMs were tagged and parsed into a structured dataset, using the `panda` library. To calculate the semantic distance between the words, the text was converted into high-dimensional numerical vectors (called “Embeddings”), using the `Sentence-Transformers` library and the `all-MiniLM-L6-v2` model. This specific model was chosen because it maps text into a dense vector space and is fine tuned to capture the semantic meaning of full sentences. This made it ideal for stand-up setups and punchlines.

Once the setups (S) and punchlines (P) were converted into vectors, Cosine Similarity was calculated to measure the angle (θ) between the two vectors, using $\cos(\theta) = \frac{S \cdot P}{\|S\| \|P\|}$.

Because comedic theory relies on separation (incongruity) rather than closeness, this similarity score was converted into the Cosine Distance (Semantic Distance) through $Distance = 1 - \cos(\theta)$.

Measuring the Semantic Distance

Recent breakthroughs in computational linguistics have mapped humour appreciation to semantic distance. (Kao, Levy and Goodman 2016) demonstrated that a joke must possess high distinctiveness between its premise and punchline to be funny, otherwise it reads as a simple statement of fact. Furthermore, (Orwig et al. 2025) proved that human appreciation for creativity and humour follows an inverted-U curve in relation to semantic distance.

If the distance is too low, the text is predictable. If it is too high, the text is perceived as a non-sequitur or hallucination.

Based on this framework, the output distances were mapped to four theoretical zones to evaluate the LLM’s performance. These boundaries are designed to reflect the Incongruity-Resolution Theory of humour, which states that a successful joke must first violate expectations (establishing incongruity) and then provide a hidden logical rule that makes sense of the surprise (resolution). The thresholds for these zones were decided by mapping this cognitive process to the mathematical properties of vector embeddings. At lower distances, the semantic overlap is too dense for the brain to register an expectation violation, resulting in straightforward cognitive understanding rather than humour. As semantic distance increases, humour builds toward the state where incongruity is maximised but remains successfully resolvable by the audience. This must fall within the higher end of the Semantic Distance, but not to so high that the joke is non-nonsensical. Beyond the upper threshold, the semantic gap becomes too vast and the cognitive effort required to find a connecting rule fails, and the brain rejects the text as illogical. These four zones provide a systematic metric to evaluate how well the model navigates the delicate psychological balance required for effective stand-up comedy.

The “Too Predictable” Zone (< 0.35)

When the semantic distance between the setup and punchline is too low, the AI fails to generate the distinctiveness required for humour, instead producing logical, conversational continuations.

Output 49	Model: GPT-3.5 ID: GxDx02
Setup: “Facebook’s where you go to see your aunt share conspiracy theories. Last week, my aunt shared a post claiming that Wi-Fi causes allergies.” Punchline: “I was like, ‘Aunt Karen, you didn’t have allergies before Facebook. Maybe it’s the stress of reading all those articles?’ ”	

Semantic Distance: 0.3368

Description: In Experiment GxDx02, the model explores the friction of generational social media usage, specifically targeting the spread of misinformation within family dynamics and using a grounded, low-semantic-distance punchline.

This falls into the “Too Predictable” zone. The vector embeddings of the punchline are mathematically too similar to the setup. Because there is no cognitive leap or misdirection, the audience experiences no ‘surprise’ factor. It reads as a plausible anecdotal statement rather than a joke.

The “Light Observation” Zone (0.35 to 0.50)

In this zone, the AI introduces a slight semantic pivot. The punchline relies on a very obvious logical leap or mild exaggeration. It could elicit only a mild reaction because the underlying mathematical connection remains very direct and safe.

Output 50	Model: GPT-3.5 ID: GxDx02
Setup: “You ever notice how your phone has become your best friend? I’ve had more meaningful conversations with Siri than I’ve had on actual dates!”	
Punchline: “And Siri never judges me when I ask her weird questions at 3 AM.”	
Semantic Distance: 0.4475	
Description: In Experiment GxDx02, the model uses a comparison between technology and human relationships, emphasising the reliability and non-judgmental nature of AI assistants compared to the social pressures of dating.	

This represents a safe, mild observation. The AI successfully creates a pivot, shifting from meaningful conversations on dates to asking weird questions

at 3 AM but the semantic domain of using a phone assistant is still very closely tied to the setup. It lacks the distinctiveness needed for a strong comedic punch.

The “Sweet Spot” (0.50 to 0.75)

This is theoretically where strong stand-up comedy lives. When the model successfully achieves optimal incongruity, the punchline introduces a completely new concept or context, creating a strong twist. Crucially, however, it remains mathematically anchored enough to the setup for the brain to instantly resolve the connection, creating the necessary “Aha!” moment.

Output 51	Model: GPT-3.5 ID: GxDx01
Setup: “And the questions! So, what do you do for fun?” Uh, I scroll through Netflix trying to find something to watch until I fall asleep. But you can’t say that, so you lie: Oh, I love reading.’ ” Punchline: “I’ve been on the same page of a book for six months.” Semantic Distance: 0.6934 Description: In Experiment GxDx01, the model contrasts the social expectation of being well-read with the reality of digital distractions, using a higher semantic distance to surprise the audience with the extent of the procrastination.	

This sits perfectly in the optimal zone. The punchline forcefully shifts the semantic context from lying on a dating app to the specific, distinct imagery of staring at a single book page for months. It creates a high degree of distinctiveness while retaining the hidden logical rule that connects the vectors back to the setup.

The “Hallucination” Zone (> 0.75)

At the other end of the spectrum, the semantic gap becomes too wide. The

AI attempts misdirection but fails to provide a logical resolution, resulting in total nonsense or absurd non-sequiturs. The vectors share almost no underlying mathematical relationship.

Output 52	Model: GPT-3.5 ID: GxDx02
Setup: “I remember when the biggest tech problem was untangling your headphones. Now, it’s like, ‘Oh no, I accidentally sent my boss a gif of a cat twerking!’ ” Punchline: “Try explaining why you look like a bunny rabbit in your performance review!” Semantic Distance: 0.8149 Description: In Experiment GxDx02, the model contrasts the physical inconveniences of older technology with the high-stakes social risks of modern digital communication, using a significant semantic distance to bridge the gap between simple technical malfunctions and professional humiliation.	

This extreme distance indicates a complete breakdown in semantic resolution. The AI introduces a `bunny rabbit` concept that shares virtually no mathematical or logical relationship with the `cat twerking` setup. The incongruity is massive, but the cognitive resolution fails, leaving the audience confused rather than amused.

Results of the Semantic Distance Evaluation

The average semantic distance was calculated for each experiment and tracked across Iterations 1, 2, and 3 (detailed in Table 5) to empirically observe if iterative prompting pushed the AI toward the optimal 0.50 – 0.75 “Sweet Spot”. This mean represents the average across all the individual jokes in the routine.

The tracking of these semantic zones across the feedback loop revealed a lot about the development of the stand-up comedy routines. Overall, there was

Model	Iter. 1	Iter. 2	Iter. 3	$1 \rightarrow 3$
GPT-4o	0.71408	0.71902	0.74096	0.02688
GPT-3.5	0.66928	0.69234	0.67164	0.00236
Google Gemini 2	0.77716	0.73648	0.74544	-0.03172
Mistral 2 Large	0.7752	0.70098	0.62732	-0.14788
Claude 3 Sonnet	0.75966	0.72192	0.78944	0.02978

Table 5: The average Semantic Distance by each LLM model across each iterations.

an observable effect of ‘reining it in’ where the models frequently started with generating more obscure premises and adjusted towards more plausible semantic connections.

The routines demonstrated in the first iteration were heavily populated by “hallucinated” (> 0.75 semantic distance) jokes. In the stand-up world, this might be called a non-sequiter, a statement that does not logically follow from the previous statement or premise. However, in Iteration 2, the hallucination rate plummeted and the volume of jokes sitting in the “Sweet Spot” substantially increased. By looking at a couple of specific examples from routines, we can see this progress happening.

Experiment GxDx02

In this example, the model “progressed” the routine by moving up at the zones towards the “Sweet Spot”. In its initial draft, the AI produced safe, borderline ‘dad jokes’, such as comparing phone usage to calling a mother (Semantic Distance: 0.4253). By Iteration 3, the prompt successfully forced the model to take calculated semantic risks. It developed a routine regarding deep conversations with Siri, ending with the punchline: "She replied, ‘I found four places nearby that sell pizza.’ Thanks, Siri, you get me!" This pushed the joke up to a distance of 0.6297, landing it perfectly within the “Sweet Spot” range.

Experiment GxDx19

Sometimes the models had to pull back a routine into a more semantically

connected state. GxDx19’s generation produced highly disjointed material, averaging an extreme distance of 0.8515. One initial joke combined dating profiles with arguing with customer service and `eating gas station sushi` (Distance: 0.8723). Through the iterative ‘feedback loop’, the prompt acted as a computational editor, forcing the model to rein in the absurdity. By Iteration 3, the model grounded its distinctiveness in logic, producing a joke about dating app photos: `Are you the one with the blurry face? Oh no, you’re the one holding the dead fish! leading to the punchline Swipe right for the halibut... and left for the catfish.` This semantic pun mathematically resolved perfectly, dropping the distance into the “Sweet Spot” at 0.5279.

These results demonstrated that the initial generation is often highly disjointed, the semantic gap being too wide or not wide enough. The application of the iterative refinement acted as an editor to rein in these semantic connections towards the “Sweet Spot” zone. A majority of the generations ended up falling within this zone by Iteration 3, and this can be seen in Table 6 per model, and in Table 7 across all the generations.

Model	Avg. Iter. 3	Zone
GPT-4o	0.74096	The “Sweet Spot”
GPT-3.5	0.67164	The “Sweet Spot”
Google Gemini 2	0.74544	The “Sweet Spot”
Mistral 2 Large	0.62732	The “Sweet Spot”
Claude 3 Sonnet	0.78944	The “Hallucination” Zone

Table 6: The average semantic distance in Iteration 3 and categories by LLM model.

Figures 20 and 21 show where the generations and LLM outputs fall into the above-mentioned categories, such as `The “Sweet Spot”` and `The “Hallucination” Zone`.

As shown in table 7, upon an analysis of each Iteration 1 routines’ jokes, 12 of the 25 generations fell into the “Sweet Spot” Zone and the remaining 13 fell

ID	Iter. 1	Iter. 2	Iter. 3	Start	End
GxDx01	0.7152	0.6884	0.7967	The “Sweet Spot”	The “Hallucination” Zone
GxDx02	0.5955	0.7332	0.6989	The “Sweet Spot”	The “Sweet Spot”
GxDx03	0.7524	0.6410	0.7122	The “Hallucination” Zone	The “Sweet Spot”
GxDx04	0.7856	0.7089	0.7190	The “Hallucination” Zone	The “Sweet Spot”
GxDx05	0.7217	0.8236	0.7780	The “Sweet Spot”	The “Hallucination” Zone
GxDx06	0.7844	0.7823	0.6977	The “Hallucination” Zone	The “Sweet Spot”
GxDx07	0.5893	0.7370	0.6936	The “Sweet Spot”	The “Sweet Spot”
GxDx08	0.6557	0.6318	0.6161	The “Sweet Spot”	The “Sweet Spot”
GxDx09	0.7246	0.7057	0.7007	The “Sweet Spot”	The “Sweet Spot”
GxDx10	0.5924	0.6049	0.6501	The “Sweet Spot”	The “Sweet Spot”
GxDx11	0.8143	0.7283	0.8850	The “Hallucination” Zone	The “Hallucination” Zone
GxDx12	0.7208	0.6364	0.7380	The “Sweet Spot”	The “Sweet Spot”
GxDx13	0.7888	0.7760	0.7202	The “Hallucination” Zone	The “Sweet Spot”
GxDx14	0.7606	0.8012	0.7144	The “Hallucination” Zone	The “Sweet Spot”
GxDx15	0.8013	0.7405	0.6696	The “Hallucination” Zone	The “Sweet Spot”
GxDx16	0.8444	0.5230	0.6467	The “Hallucination” Zone	The “Sweet Spot”
GxDx17	0.7724	0.7935	0.6955	The “Hallucination” Zone	The “Sweet Spot”
GxDx18	0.6938	0.7972	0.5875	The “Sweet Spot”	The “Sweet Spot”
GxDx19	0.8515	0.6503	0.5943	The “Hallucination” Zone	The “Sweet Spot”
GxDx20	0.7139	0.7409	0.6126	The “Sweet Spot”	The “Sweet Spot”
GxDx21	0.7572	0.6954	0.7657	The “Hallucination” Zone	The “Hallucination” Zone
GxDx22	0.6452	0.6766	0.8089	The “Sweet Spot”	The “Hallucination” Zone
GxDx23	0.8497	0.7681	0.7965	The “Hallucination” Zone	The “Hallucination” Zone
GxDx24	0.7380	0.6987	0.7676	The “Sweet Spot”	The “Hallucination” Zone
GxDx25	0.8082	0.7708	0.8085	The “Hallucination” Zone	The “Hallucination” Zone
Count				12	17
%				48%	68%

Table 7: Comparison of average Semantic Distance and categories for each experiment.

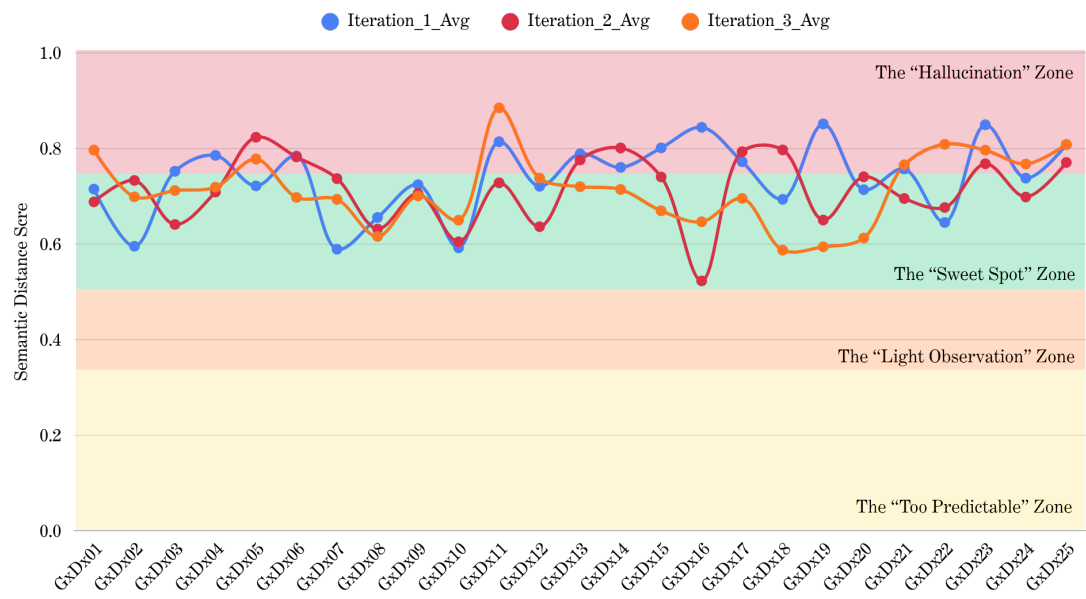


Figure 20: A graph showing the Semantic Distance across all the generations, with category zones

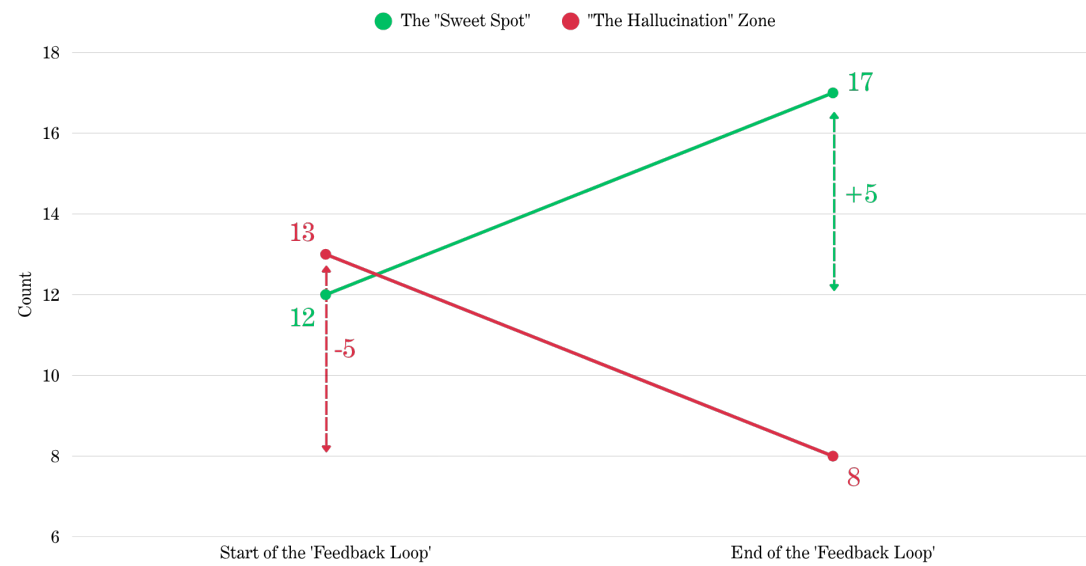


Figure 21: A graph showing the count of each category of Semantic Distance scores

into the “Hallucination” zone. Whereas in Iteration 3, 17 of the 25 generations fell into the “Sweet Spot” zone and only 8 fell into the “Hallucination” zone, as shown in Figure 21. Overall, as shown in Table 6, 4 out of the 5 models average fell into the “Sweet Spot” after Iteration 3, with only Claude 3 Sonnet being in the “Hallucination” zone. This mathematical shift provides critical evidence of the “Progression and Development” component of creativity in computational systems Jordanous and Keller (2016). In the context of creative generation, “progression” requires a system to move beyond its initial, raw output and actively refine its artifacts toward a domain-specific goal. The data proves that the models achieved this.

The AI did not remain trapped in the “Too Predictable” (< 0.35) or “Hallucination” (> 0.75) zones. It used the feedback loop to recognise its current state and actively adjust its semantic choices by learning to balance high distinctiveness (the surprise) with a mathematically sound connection (the resolution).

The observable trajectory of jokes moving from a 0.85 distance down to a 0.59 distance over three iterations is direct, mathematical proof of “development”. The vector analysis confirms that the LLMs are not merely generating static text. Through iterative prompting, they are actively developing a comedic concept. This process of moulding disconnected semantic vectors into a cohesive, optimal joke demonstrates a clear, measurable progression of the creative artifact, proving that the iterative feedback loop is essential for developing domain-appropriate stand-up comedy.

In Figure 22 all of the individual jokes have been scored by semantic distance and categorised accordingly, showing that a majority of the jokes fall into the “Sweet Spot” and “Hallucination” zones. These zones being the most likely level of incongruity to generate humour.

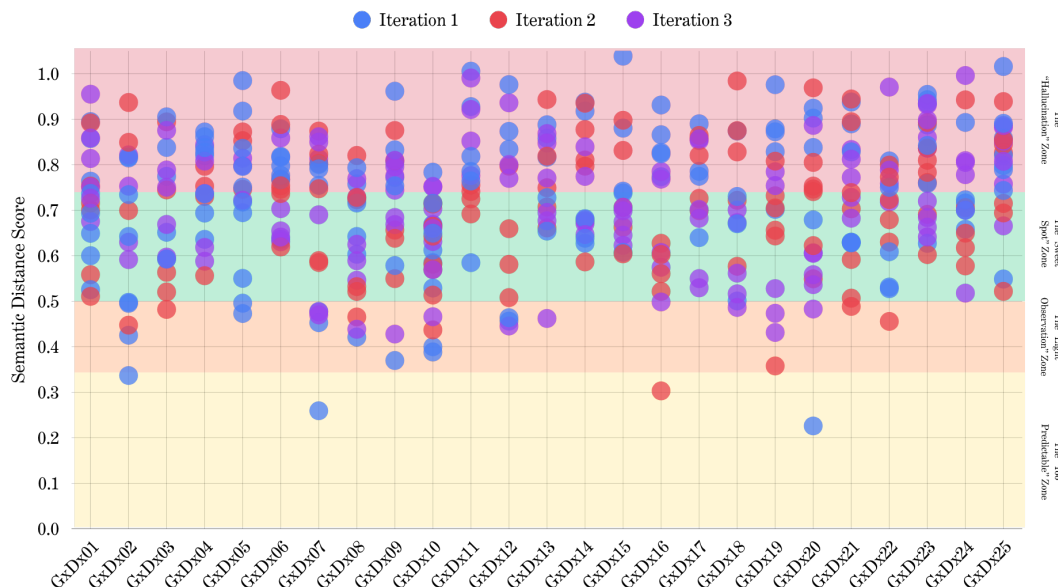


Figure 22: A graph showing the the Semantic Distance scores of all individual jokes from the dataset, with zones

5.4.4 Stand-Up Techniques and Improvements

Many of the generations relied on broad, universally recognised tropes and cliches. In initial iterations, the LLM models frequently defaulted to well-covered observational concepts. In many examples of feedback during the reflection/review stage, the suggestion was for **More Personal Stories** and **Avoid Cliches**. This is useful feedback because it leads to the routine having more novelty, rather than typicality. In these examples, the models showed clear progression through actioning this feedback. This is evidenced through specific examples throughout the dataset.

Moving Beyond Cliche's

Generation GxDx01

In GxDx01, the original text reads, "Swipe left, swipe right - it's like

shopping for humans on Amazon." This is a well-worn trope. Following feedback to sharpen punchlines and vary topic depth, the LLM revised this to: "Tinder has turned us into human slot machines. 'Jackpot! I matched with someone who's 90% less creepy than the last one!'" The progression here is evident: the model shifts from a passive observation to an active, gamified metaphor complete with hypothetical dialogue from an undefined character.

Generation GxDx04

In GxDx04, the initial joke relied on the overused visual of "business up top, pyjama party down below." Prompted for "Additional Characters," the model introduced a highly specific familial intrusion: "Like when your kid walks in... dressed as Spider-Man... 'Sorry, team. Gotta go. Spider-Man needs his PB&J.'" This unnamed 'kid' character is the improvement that was recommended within the feedback transforming a generic Zoom observation into a vivid, relatable scenario.

Generation GxDx02

Similarly, in GxDx02, the LLM originally produced a highly typical tech joke: "I remember when the biggest tech problem was untangling your headphones." After feedback requested a Stronger Opening Joke and to Avoid Clichés, the model shifted to a much more personalised, slightly tragicomic premise: "I've had more meaningful conversations with Siri than I've had on actual dates! And Siri never judges me when I ask her weird questions at 3 AM." This demonstrates a leap from a generic nostalgic observation to character-driven self-deprecation.

Subverting Expectations and Absurdism

As the LLM processed deeper feedback, it began to demonstrate an ability to subvert audience expectations, moving away from predictable logical conclusions

toward productive absurdity.

Generation GxDx03

In GxDx03, the initial attempt at family-based humour was highly stereotypical: "Uncle Bob... shows up at every family gathering in a Hawaiian shirt." Prompted to include *Variety in Jokes* and "Subverting Expectations," the LLM escalated the imagery significantly: "Nothing says 'rest in peace' like a shirt covered in flamingos. I mean, who dressed this guy? Jimmy Buffet?" By introducing the jarring context of a funeral ("rest in peace") alongside the flamingo shirt, the model successfully subverts the typical "wacky uncle at a BBQ" trope.

Generation GxDx06 and GxDx07

The GPT-3.5 LLM also showed an improved capacity to extend metaphors into unexpected territory. In GxDx06, comparing online dating to "shopping for groceries on an empty stomach" was revised into a deeper, more absurd visual: "...you're left wondering why you're holding a can of beans with no label." In GxDx07, a simple mention of a profile picture with a llama was elevated to: "It's like showing up to a magic show and realizing the magician's biggest trick is making your interest disappear."

Structural Changes and Punchlines

"Progression" in comedic writing is often measured by the economy of words and the impact of the punchline. The dataset reveals that the LLM successfully provided and actioned structural feedback, such as applying the Rule of Three and punchline tightening.

Generation GxDx14

In GxDx14, the initial premise was weak: "The only thing flatter than

an ironed shirt is my social life after 8 pm." After being prompted to tighten the punchline, the LLM generated: "...it comes out looking like a deflated whoopee cushion with a serious case of commitment issues." The revised imagery is far more evocative and absurd, showcasing a sophisticated layering of physical comedy (whoopee cushion) with emotional anthropomorphism (commitment issues).

Generation GxDx13

In GxDx13, the mundane observation that a fridge "becomes a crime scene" was pushed to deepen the joke. The LLM responded by building tension and a more vivid micro-world: "You open the door... and BAM! Staring you down is the Tupperware graveyard..."

Generation GxDx12 and GxDx22

This structural tightening is also evident in GxDx12, where an overly wordy joke about adulting ("basically like that gym membership you never use... you get hit with a random annual fee") was condensed into a sharper, punchier rhythm: "Adulting is basically a never-ending game of whack-a-responsibility." In GxDx22, a generic complaint about self-checkout machines was tightened to include a narrative payoff: "As I'm leaving, feeling defeated, I spot the store manager struggling with the same machine. Turns out, even the pros can't handle these silicon monsters."

The model shows an understanding that a punchline must escalate the premise rather than merely restate it, indicating a progressive grasp of these comedic mechanics. It simultaneously demonstrates an understanding of the "tightening" process in stand-up that was discussed in Section 2.5.2 of the Background chapter.

The Performance

A significant marker of “development” in the LLM’s outputs is its transition from writing text meant to be read to scripting material meant to be performed. Although the LLM cannot perform the routine itself, it is able to add content that will enhance the stage performance.

Generation GxDx18, GxDx09 & GxDx05

Feedback frequently requested `Physical Comedy` or `Act-Outs`. In GxDx18, the original line (`"brighter than my future"`) was entirely verbal. The revision introduced self-deprecation tied to physical identity and stage directions: `"Maybe it's because I'm a ginger... [Demonstrates a silly moonwalk]."` Similarly, in GxDx09, the LLM includes `"[Comedian pulls out phone and pretends to swipe]."` In GxDx05, an observation about IKEA furniture is elevated by the bracketed direction: `"[pretends to struggle assembling furniture],"` proving the model understands when a physical act-out serves as the punchline itself.

Generation GxDx24 and GxDx21

The LLM learned to utilise the audience as an active component of the routine. In GxDx24, the abstract thought about parents `"winging it"` was grounded into direct crowd work: `"Quick poll: who here has successfully kept a plant alive for more than a month?"` Furthermore, in GxDx21, an opening line stating `"I've been trying this whole online dating thing"` was transformed into dynamic interaction: `"Who here's tried online dating? ... Wow, I feel less alone already... and more concerned for humanity."` Here, the LLM scripts a built-in pause for audience response, using their hypothetical reaction as the setup for the secondary punchline.

Callbacks and Tags

Generation GxDx17 and GxDx25

Finally, the LLM demonstrated an emerging ability to sustain comedic narratives through tags (additional punchlines following the main joke) and callbacks (referencing earlier material). In GxDx17, a joke about bagged milk is extended with a tag: "At least it's not as bad as the ice cream soup waiting in my car," maintaining the comedic momentum. In GxDx25, a joke about faking connection issues on a work call is given a much stronger, narrative-driven closing tag: "Then you gotta commit and throw it [your phone] in the toilet. 'Oh no, my phone accidentally fell in! Guess I can't work for a week!'"

5.4.5 Summary

This 'simulated feedback loop' remains the most important step for assessing the LLMs' ability to progress and develop their responses to a task. Through multiple iterations, significant changes were made by all models, resulting in vastly different and structurally richer stand-up comedy routines. A direct, measurable link was identified between the internal review and the material improvements made in the routines. This shows that the LLM is highly intentional in the improvements it is making, demonstrating a clear act of computational creativity.

The data shows that LLMs are capable of executing an engaged and reflective, multi-stage editorial process. This progression is evidenced across a few different measurements.

- The aggressive edit rates, measured via the Normalised Levenshtein Distance, indicate that the models do not simply append or tweak text, rather they actively reconstruct it. While this occasionally leads to structural instability

(the “thrashing” pattern), it ultimately results in profound transformations of the stand-up comedy routine draft.

- By mapping the semantic distance between setups and punchlines, this evaluation provides empirical proof of comedic development. The models successfully used the iterative feedback loop to move their outputs out of the logical “Too Predictable” zone or the nonsensical “Hallucination” zone, actively steering their semantic distinctiveness into the optimal comedic “Sweet Spot”, showing that the the models are refining its output toward the domain-specific goal of creating effective jokes.
- The models demonstrated a progressive understanding of their own feedback and how to improve themselves. Over successive iterations, they successfully shed universal clichés in favour of specific, character-driven absurdities. The intentional inclusion of stage directions, crowd work, physical comedy cues, and narrative callbacks proves that the LLM can produce contextual text not just as written prose, but as a live performance script.

The simulated feedback loop acts as a vital computational editor for working with LLMs. It proves that when guided by precise, self-generated critique, LLMs exhibit a measurable capacity for “Progression and Development”, transforming raw, predictable text generation into cohesive, intentional, and structurally sound creative artifacts.

5.5 Phase 4: Performance Annotations

The final requirement for this study was to take the refined stand-up comedy routines and purposefully add directions and instructions to the performer detailing how to deliver the material. This step addresses the ‘Text-Performance Gap’ discussed in Section 2.4.1 of the Background chapter, moving the artifact from a

purely literary text to a functional performance script by encoding ‘oral punctuation’ and paralinguistic features.

To evaluate this phase, the models were assessed on whether they met the requirements of the prompt and whether they demonstrated sufficient “Progression and Development” by expanding beyond the provided examples. This step proves a more difficult task for an LLM to show this aspect of creativity, as it has been given a specific list to use as a reference. Showing unique annotations would be a strong indicator of creativity, as it demonstrates innovation despite explicit instructions and constraints.

5.5.1 Annotation Frequency

All five models successfully applied performance annotations to their generated routines, as detailed in Table 8 and shown in Figure 23.

Model	Gen. 1	Gen. 2	Gen. 3	Gen. 4	Gen. 5
GPT-4o	15	34	17	17	29
GPT-3.5	9	18	19	17	25
Google Gemini 2	18	12	16	29	24
Mistral 2 Large	38	33	52	36	43
Claude 3 Sonnet	53	26	36	27	33

Table 8: The raw number of annotations added to each generation.

However, through a simple word count, it was clear the routines were of variable lengths. The number of annotations applied is inherently related to the length of the routine (i.e., longer routines provide more opportunities for annotations). Therefore, the raw number of annotations was measured against the word count of each routine (before annotations were applied) to generate a standardised frequency score, detailed in Table 9 and shown in Figure 24.

When adjusting for length, the data clearly shows that the Mistral 2 Large model achieved the highest average score (0.06), resulting in the most annotations

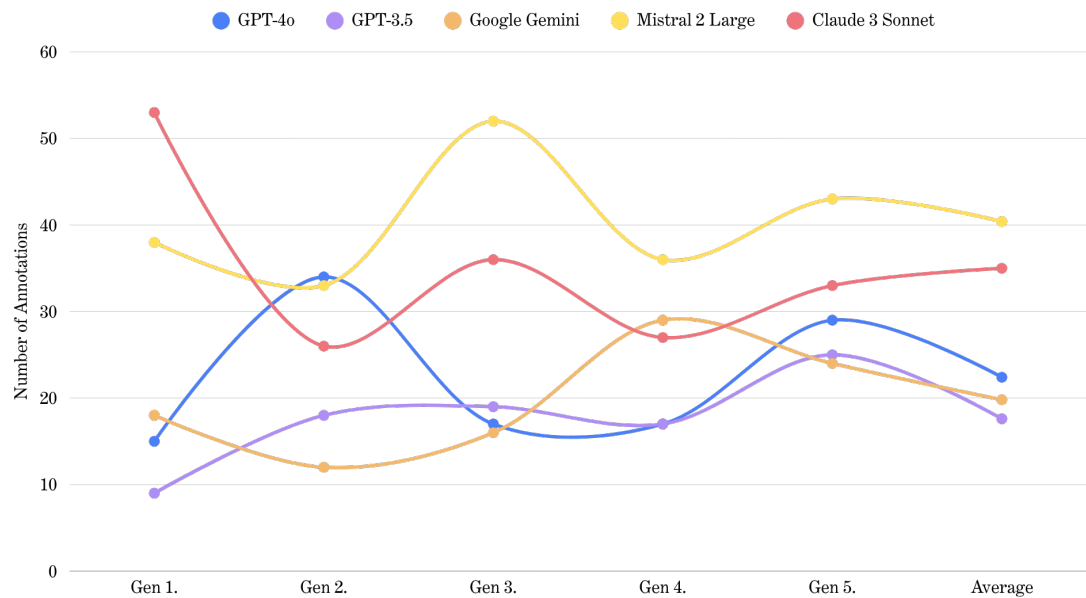


Figure 23: A graph showing number of annotations added per model

Model	Gen. 1	Gen. 2	Gen. 3	Gen. 4	Gen. 5	Sum	Avg.
GPT-4o	0.02	0.05	0.02	0.03	0.04	0.17	0.03
GPT-3.5	0.02	0.03	0.03	0.04	0.05	0.18	0.03
Google Gemini 2	0.05	0.03	0.03	0.07	0.05	0.23	0.05
Mistral 2 Large	0.05	0.05	0.08	0.08	0.06	0.32	0.06
Claude 3 Sonnet	0.07	0.04	0.05	0.05	0.05	0.27	0.05

Table 9: The normalised score based on annotations relative to word count.

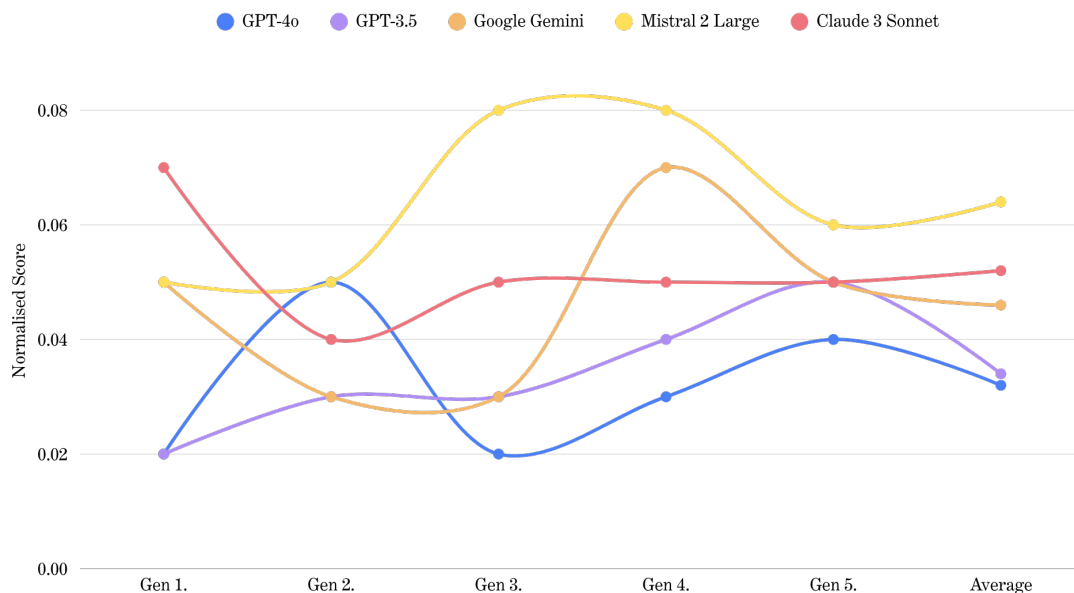


Figure 24: A graph showing number of normalised annotations added relative to routine length per model

relative to the length of the routine. Claude 3 Sonnet and Google Gemini 2 also performed well (0.05), whereas the GPT-4o and GPT-3.5 models performed far weaker in sheer volume (0.03).

5.5.2 Detail and Qualitative Scoring

While volume provides a baseline metric, the detail and style of the annotations varied significantly between models and individual generations. An investigation into the text revealed distinct stylistic choices, as shown in Table 10.

To systematically assess this variance, each generation was rated on a qualitative scale from 1 to 3 based on the detail of their annotations, shown in Table 11:

- **Score of 1:** Represents minimal annotations (e.g., Google Gemini 2, GxDx14).
- **Score of 2:** Represents standard annotations (e.g., Claude 3 Sonnet, GxDx24).

Model	Observations	Examples and Identifiers
GPT-4o	Demonstrated highly variable detail. Utilised highly detailed, consistent structures in some instances, while opting for minimal directions in others.	GxDx01: (Volume: Loud, Hand Gestures: Thumbs up) (Facial Expressions: Raised Eyebrow, Volume: Low) GxDx04: (Shrugs, raised eyebrow) (finger wag, frown, quick pacing)
GPT-3.5	Applied a very consistent structure using square brackets. Distinguished clearly between spoken material delivery and specific laugh breaks.	GxDx07: [Emphasis on “Hugging”, wide arm spread] [Pause for laughter, smiling with a cheeky grin]
Google Gemini 2	Tended toward minimal, useful notes in simple brackets. Focused on basic delivery but lacked structural consistency.	GxDx14: (Normal Pace), Wrinkles nose, Frown
Mistral 2 Large	Applied annotations similarly to Gemini 2 but with significantly higher information density, comparable to GPT-4o’s level of detail.	GxDx17: (Smile, look around, Stage Presence: Active, Movement: Walk Around) (Pause: Short, lean in, act out the scenario, Hand Gestures: Pointing, Stage Presence: Tone Lean in for intimacy)
Claude 3 Sonnet	Produced a large volume of annotations characterized by clear, categorised detail.	GxDx24: [Delivery – Pace: Speed Up] [Stage Presence - Movement: Active]

Table 10: Comparative analysis of LLM model’s annotation styles and structural consistency.

- **Score of 3:** Represents highly detailed annotations (e.g., GPT-4o, GxDx01).

Model	Avg. Score	No. of 3 Scores
GPT-4o	2.2	2
GPT-3.5	2.8	4
Google Gemini 2	1.4	0
Mistral 2 Large	2.2	3
Claude 3 Sonnet	1.4	0

Table 11: The average annotation detail score for each model and the count of max scores.

If we apply descriptive threshold categories to this (Below 1.5 = Below Expectations; 1.5 to 1.9 = Expected; 2.0 or above = Beyond Expectations), the results show clear performance tiers, visible in 12:

Model	Category
GPT-4o	Beyond Expectations
GPT-3.5	Beyond Expectations
Google Gemini 2	Below Expectations
Mistral 2 Large	Beyond Expectations
Claude 3 Sonnet	Below Expectations

Table 12: Categories based on each model’s average detail score.

According to these qualitative results, the best-performing model was GPT-3.5, which achieved the highest average score of 2.8 and produced the most top-rated generations. GPT-4o and Mistral 2 Large also performed strongly, successfully developing the response past the basic instructions in the prompt. Conversely, Google Gemini 2 and Claude 3 Sonnet averaged scores of 1.4, placing them “Below Expectations” regarding qualitative detail.

5.5.3 Progression Beyond the Prompt (Uniqueness)

To prove that the models were acting creatively rather than engaging in mere generation (Ventura 2016), the annotations were evaluated for uniqueness. This

measured how many annotations were uniquely written by the Large Language Models, meaning they were ‘inspired’ or “developed” from the prompt but did not explicitly exist in the provided sample list.

General performative actions not found in the prompt examples were counted as ‘new’ annotations. For example, Claude 3 Sonnet (GxDx24) generated the examples detailed below.

Output 53	Model: Claude 3 Sonnet ID: GxDx24
• Facial Expression: raised eyebrow with a slight head tilt <i>[Head tilts were not in the sample provided to the LLM]</i> • Facial Expression: wide eyes, slight head nod <i>[Head nods were not in the sample provided to the LLM]</i> Description: The Claude 3 Sonnet model exhibited emergent behavior by generating stage directions for physical gestures that were not present in the original training sample, such as head tilts and nods, suggesting an internal mapping of comedic timing and physicality.	

Routine-specific annotations were excluded from this calculation, as they are entirely contextual to the joke rather than general performative markers that could be applied across different routines. For example, Claude 3 Sonnet (GxDx24) also generated `acts out rolling down window`. This was deemed too specific to qualify as evidence of generalised annotation development.

The percentage of annotations that demonstrated “development” is detailed in Table 13 and shown in Figure 25

Across the dataset, 88% of the generations demonstrated “progression and development” beyond the samples given in the prompt. The models acted creatively by using the provided samples as a starting inspiration and applying entirely new

Model	Gen. 1	Gen. 2	Gen. 3	Gen. 4	Gen. 5	Avg.
GPT-4o	0%	8.82%	17.65%	17.65%	10.34%	10.89%
GPT-3.5	0%	22.22%	36.84%	5.88%	0%	12.99%
Google Gemini 2	16.67%	25.00%	18.75%	13.79%	8.33%	16.50%
Mistral 2 Large	18.42%	6.06%	5.77%	5.56%	4.65%	8.09%
Claude 3 Sonnet	7.54%	19.23%	5.55%	14.81%	12.12%	11.85%

Table 13: The percentage of ‘inspired’ or ‘developed’ unique annotations in each generation.

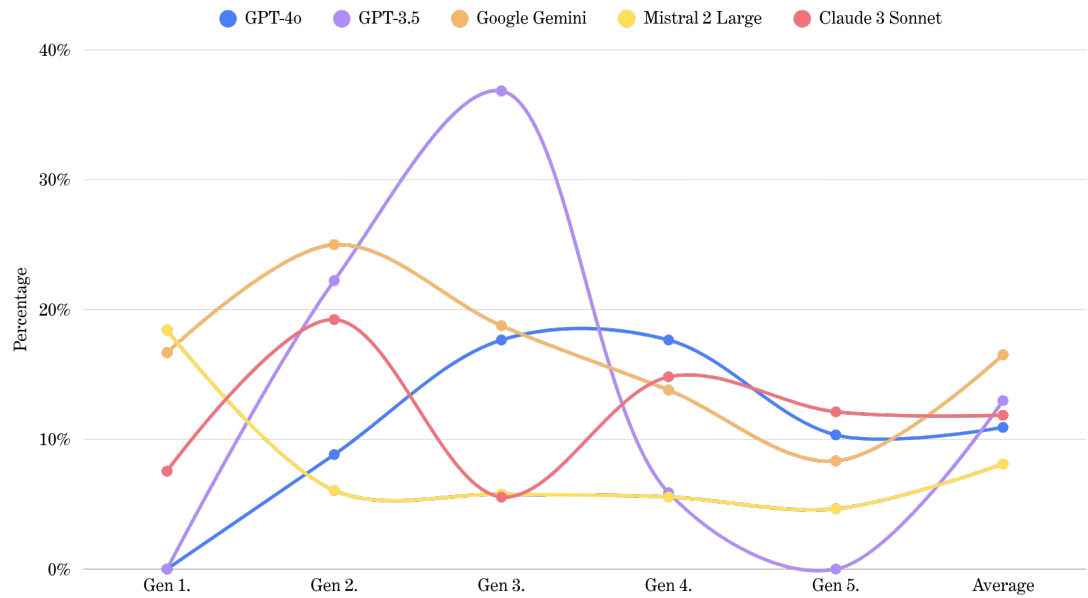


Figure 25: A graph showing percentage of annotations added that weren’t specified in the prompt

annotation ideas in suitable contexts.

The Google Gemini 2 model performed the strongest in generating unique annotations (averaging 16.50%), while the Mistral 2 Large model performed the weakest (8.09%). Interestingly, this is the exact reverse of their performance in annotation volume, where Gemini 2 generated the fewest and Mistral 2 generated the most. By successfully interpreting a sample list and generating new, contextually appropriate performance markers, the models proved they could partially bridge the ‘text-performance gap’ and push their outputs beyond standard expectations.

5.6 Survey and External Review

5.6.1 Design

While Phase 3 of the evaluation successfully measured the internal progression of the large language models through metrics like semantic distance, edit rates and comedic decisions, an audience survey was designed to provide external insight and measure human reception, in order to validate whether the “progression and development” achieved by the models’ internal auditor translates to actual comedic improvement,

The experiment consisted of an anonymous survey where participants evaluated LLM-generated routines using a blind A/B/C testing methodology, as detailed in Figure 26. Participants were presented with 10 randomly ordered trios of jokes. Each trio featured the same or similar joke extracted from the different stages in the ‘feedback loop’: Iteration 1, Iteration 2 and Iteration 3. The joke trios were sampled at random from a list of all jokes across the routines. The participants were tasked to pick the best or funniest option from each trio. The survey was presented to the general public to complete anonymously, to emulate the concept of a ‘walk-in’ stand-up comedy gig in which anybody could walk in and hear the jokes.

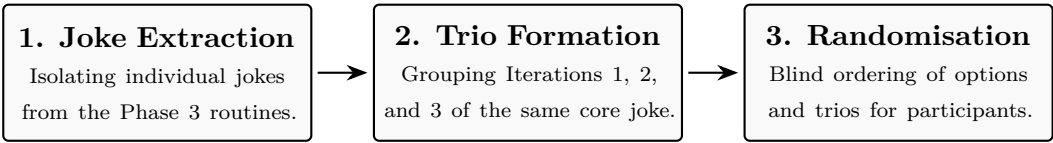


Figure 26: Methodology of extracting jokes for the external survey

Table 14 shows an example trio from the survey, extracted from the Claude 3 Sonnet generation GxDx22:

Stage	Description	Example Extract
Iteration 1	The initial draft of the generated material	I see it. The self-checkout area. It's like the Bermuda Triangle of retail. Items go in, sanity disappears
Iteration 2	The revision	It's 2 AM, I'm at the 24-hour supermarket in my pajamas. It's me versus the machine. Like "The Terminator," but with more beeping and less Arnold. Ever feel like these machines are judging your life choices?
Iteration 3	The final targeted draft	I used to think I was pretty tech-savvy. I can doom-scroll on Twitter for hours, but nothing has humbled me quite like the self-checkout machine. It's like the ninth circle of retail hell.

Table 14: Categorisation and examples of the generated material across iterative stages.

By randomising the order of the jokes and the trios themselves, the experiment ensured a blind test of the material’s humor and structural quality. Participants have no idea which joke of the three is Iteration 1, 2 or 3.

Participants were presented with a mandatory first wave of 10 joke trios (with an optional second wave of an additional 10 trios). Following the quantitative selection, participants were invited to provide qualitative comments detailing their reasoning, their perception of the jokes’ structures, and any specific flaws they

noticed.

The earlier analysis utilising the Levenshtein distance in Section 5.4.1, qualitative analysis in Section 5.4.4 and Semantic Distance in Section 5.4.3 indicated that the iteration 3 performed better internally and showed clear development. The hypothesis would be that the LLM’s self-reflection loop would successfully map to human comedic standards.

5.6.2 Voting Distribution

The aggregate data collected from the participants answers revealed a surprising trend that contradicted the internal metrics. Rather than preferring the iterated latter routines, many participants favoured the earlier drafts. Iteration 1 received the majority of votes at 37.7%, closely followed by Iteration 2 at 36.7%. The final drafts (Iteration 3) performed a little less, securing only 25.6% of the overall jokes. This is shown in Table 15.

Stage	Audience Preference (%)
Iteration 1	37.7%
Iteration 2	36.7%
Iteration 3	25.6%

Table 15: Audience preference voting percentages across each iteration stage.

The results indicate that there is a disconnect between the LLM’s internal optimisation metric and human comedic appreciation. While iterative prompting successfully pushed the model to refine its output, performing well in computational metrics as seen in Sections 5.4.1, 5.4.3 and 5.4.4, this process was often detrimental to the subjective quality of the humour.

By Iteration 3, the model’s attempts to ‘develop’ the routine was over-corrected making it perform worse in the public vote. Human participants seemed to prefer the rawer, less filtered outputs of Iteration 1 and Iteration 2 (with there being only a 1% difference in the proportion of votes between the two).

5.6.3 Distribution (Waves)

An optional aspect of the survey was a second wave of jokes, that participants could choose to (or not) do. This provides this research with more data, as this second waves contained another 10 random samples of joke trios from the dataset, as well as gaining more data from the participants who volunteered to give more of their time to the survey.

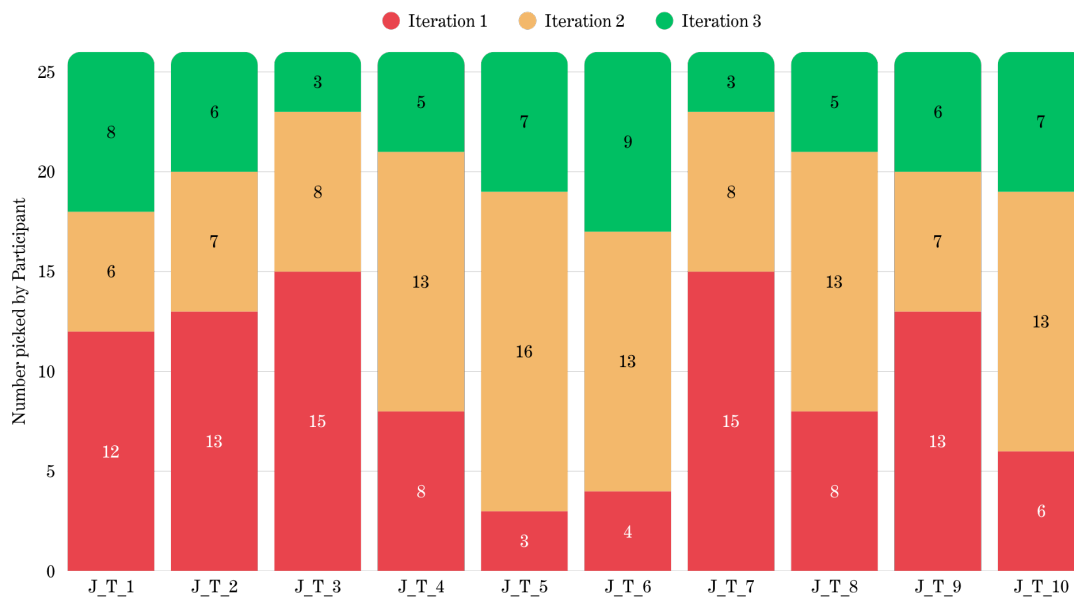


Figure 27: A graph showing the number of votes of each iteration, across wave 1

Wave 1:

During the initial, mandatory phase of 10 joke trios, audience preference peaked for Iteration 1, which captured 40% of the vote, as shown in Figure 27.

Wave 2:

For participants who opted to complete the second wave of joke evaluations, the preferences balanced to each iteration gaining a similar proportion of votes. In the second wave, Iteration 2 performed the worst, as shown in Figure 28. The results from both waves together are compared in Figure 29.

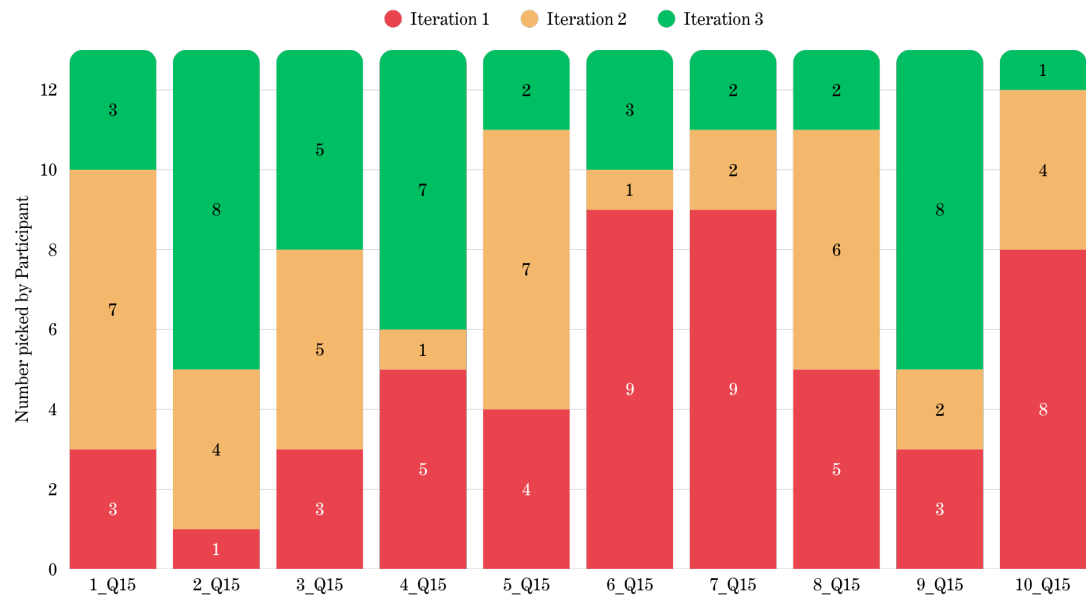


Figure 28: A graph showing the number of votes of each iteration, across wave 2

Stage	Audience Preference (%)
Iteration 1	38.46%
Iteration 2	30.00%
Iteration 3	31.54%

Table 16: Wave 2: Audience preference voting percentages across each iteration stage.



Figure 29: A graph showing the voting percentage of each iteration, across the two waves

The quantitative results from Wave 1 contradicts the primary hypothesis. The data seems to demonstrate an inverse relationship between the LLM’s ‘refinement’ and the audience preference. However, in the second wave the trend seems to change substantially. As the second wave contained a different set of jokes, but with the same participants, it shows the subjective nature of humour and how participants can enjoy some jokes but not others. This pattern allowed us to split the participants into two groups, those more interested in the experiment and those less interested in the experiment. Those more interested completed both rounds of questioning and often provided written comments.

5.6.4 Performance by Model

A breakdown of the survey by individual large language models reveals that the data varies depending on the model that generated the stand-up comedy material. When analysing the performance on a per-model basis, there is compelling evidence

that, for many models, the iterative ‘feedback loop’ frequently achieved its goal of successfully translating internal refinement into human-preferred comedic outputs by the final iteration, as shown in Table 17 and 18

Model	Iteration 1	Iteration 2	Iteration 3
GPT-4o	48.08%	25.00%	26.92%
GPT-3.5	44.23%	40.38%	15.38%
Google Gemini 2	13.46%	55.77%	30.77%
Mistral Large 2	44.23%	40.38%	15.38%
Claude 3 Sonnet	36.54%	38.46%	25.00%

Table 17: Wave 1: Audience preference voting percentages broken down by individual model across each iteration stage

Model	Iteration 1	Iteration 2	Iteration 3
GPT-4o	15.38%	42.31%	42.31%
GPT-3.5	30.77%	23.08%	46.15%
Google Gemini 2	50.00%	30.77%	19.23%
Mistral Large 2	53.85%	30.77%	15.38%
Claude 3 Sonnet	42.31%	23.08%	34.62%

Table 18: Wave 2: Audience preference voting percentages broken down by individual model across each iteration stage

The preference for Iteration 3 jokes is most prominently displayed in the second wave of testing. In this wave, several models demonstrated that their final drafts resonated more strongly with the audience.

GPT-3.5

This model showed the most success through the ‘feedback loop’. In Wave 2, Iteration 3 emerged as the clear audience favourite, capturing 46.15% of the votes from participants. This significantly outperformed the initial draft (Iteration 1) at 30.77% and the first revision (Iteration 2) at 23.08%.

GPT-4o

This model also displayed a substantial alignment with human preferences in its final iteration during Wave 2. Iteration 3 secured 42.31% of audience preferences, representing a massive improvement over its Iteration 1 score of just 15.38%. For the jokes from this model, the final draft tied for the highest preference alongside Iteration 2 (42.31%).

Claude 3 Sonnet

This model demonstrated significant preference towards Iteration 3, with 34.62%. Although under performing compared with Iteration 1, it outperformed Iteration 2 which received 23.08% of the votes.

Even within the first wave, which had a larger participant sample size than Wave 2, specific models showed clear evidence that their final iterations were an improvement to participants over earlier drafts.

Google Gemini 2

The model's final targeted draft (Iteration 3) secured 30.77% of the audience vote. This successfully outperformed its initial Iteration 1 baseline, which received the lowest score for that model at 13.46%.

GPT-4o

Continuing a trend of improvement, GPT-4o's Iteration 3 (26.92%) managed to edge out its intermediate Iteration 2 draft (25.00%) in the initial wave.

These model-specific breakdowns challenge the conclusions drawn from the aggregate data alone. They indicate that while the 'feedback' loop can produce mixed results across different LLM architectures, it is highly capable of generating the most preferred comedic output under the right conditions. Models like GPT-3.5

and GPT-4o successfully aligned their optimisations and improvements with human comedic appreciation, providing evidence that Iteration 3 can indeed represent the best quality output of the three iterations.

5.6.5 Participant Separation

In order to better understand why the aggregate data skewed away from the final LLM-generated draft, the 26 participants were segmented based on how they interacted with the Iteration 3 choice. Analysing their selection patterns reveal three distinct behavioural zones:

- **Fluctuating**

The majority of participants exhibited unpredictable choices driven by the context of the specific joke rather than a preference for a specific iteration stage. Their selections shifted dynamically across the scale, indicating that the LLM's edits were hit-or-miss depending on the topic.

- **Active Divergence**

This group consistently and actively rejected Iteration 3 in favour of the baseline drafts. This indicates a strong logical or ideological departure from the LLM's refined comedic framing, finding the earlier iterations consistently funnier.

- **High Adherence**

A small minority appreciated the improved jokes. These participants consistently chose Iteration 3 or its immediate neighbors, indicating a stable level of agreement with the target framing.

5.6.6 Consensus and Outliers

By calculating the selections of the ‘average’ participant from the dataset, the deviation of each participant from this ‘average’ person highlights the extreme subjectivity of humour and the LLM’s struggle to capture the humour of the average person. The data identifies several “Total Individuals” (e.g., Participants 4, 7, 14, 16, 19, 20, 26) whose views differed vastly from the general consensus, acting as major outliers. Conversely, the “Follows the Crowd” segment (e.g., Participants 5, 9, 15, 17, 21, 24) showed responses almost identical to the statistical average. The LLM’s refinement process (seen in options for Iteration 3) failed to unify these groups, often alienating the consensus.

While Iteration 3 performed poorly overall with this sample group, there were clear examples of individuals who predominately preferred it. For instance, Participant 1 showed high adherence to Iteration 3, as well as Participant 20 who also highly favoured Iteration 3. However, the deviation calculation flagged them as a “Total Individual” outlier, suggesting that the LLM’s refined humor appeals to a highly specific, niche audience rather than a broad crowd.

5.6.7 Qualitative Feedback

At the end of the set of 10 jokes (questions), participants were offered the opportunity to write feedback on the jokes. Not all participants did this but those that did provide an insight into their decision-making process. It offers crucial insights into why the quantitative data skewed away from Iteration 3.

Participant 20’s qualitative feedback says a lot about the performance of Iteration 3, stating that *I was just choosing the ‘best’ world salad*. As this participant chose Iteration 3 frequently (and Iteration 2 the other times), it suggests that linguistically these jokes were better, due to the iterative refinement and performance in computational metrics, but were lacking in the humour aspects.

Several thematic failure points emerged regarding the LLM’s editing process.

Chaos

While the LLM’s internal metrics (Phase 3) viewed aggressive rewrites as positive progression, human readers found the results structurally chaotic. The editing process frequently led to disjointed narratives, prematurely abandoning premises before the punchline landed.

- Participant 4 noted, “*many are quite chaotic or disjointed. instead of funny*”.
- Participant 11 observed, “*many of them were quite confusing in that they jumped into different topics within the same joke*”.
- Participant 3 added, “*Some were disjointed, switching to a new topic without resolving the previous one*”.

UK Cultural Disconnect

The LLMs relied heavily on American-centric references that alienated international participants and ruined comedic timing.

- Participant 10 explicitly stated, “*Most of them were very US-centric and required contextual knowledge that non-Americans don’t have*”.
- Participant 6 confirmed this barrier: “*Some of them had stuff that I needed to Google to find out what they meant, such as W-2, which I realised was an American thing*”.

Cliché

Feedback indicated that as the LLMs refined the jokes, they often felt *well-repeated* or derivative. This suggests the self-reflection loop sometimes forces the LLM back to the safety of its training data rather than generating novel incongruity.

- Participant 3 remarked, *“I felt like I’d seen some before, as if they were either directly from existing material, or generally well-repeated jokes”*.
- Participant 12 similarly pointed out, *“The Brenda bag one I have definitely heard before!”*

Semantic Incoherence

In later iterations, the LLMs occasionally made logical leaps that confused rather than amused, stretching the ‘Semantic Gap’ too far.

- Participant 20 described the experience as choosing the *“‘best’ world salad,” adding that “the jokes came across as generated text. There was nothing there”*.
- Participant 3 provided a direct example of a failed semantic connection: *“Some made more sense than others (for example: “‘Have you tried turning the wifi off and on?’ What, are we trying to summon a genie?” makes little sense!)”*.

Sentiment of the Comments

Analysing the 26 written comments reveals a distinct sentiment distribution as shown in Table 19.

Sentiment	Number of Comments	Proportion (%)
Negative	16	61.5%
Neutral / Mixed	8	30.8%
Positive	2	7.7%
Total	26	100%

Table 19: Distribution of overall sentiment across the qualitative feedback comments provided by survey participants.



Figure 30: A selection of qualitative comments with highlighting applied to map the distribution of positive, negative, and neutral sentiments.

The overarching sentiment of the participant feedback skewed heavily negative regarding the comedic value of the text, as seen in Figure 30, with some neutral/mixed observations about the medium, and very few purely positive reactions.

Negative Sentiment

The majority of the comments expressed active dissatisfaction with the humour. Participants frequently used blunt phrasing to describe the material, stating it was “*not funny*”, “*mediocre at best*”, or a “*sad selection*” where “*none of these encouraged a smile.*” Many participants noted they were not voting for a joke they enjoyed, but rather choosing the “*‘best’ word salad*” or identifying the “*least funny*” option.

Neutral Sentiment

A notable portion of the feedback was neutral, focusing objectively on the mechanics of the jokes rather than the humour itself. These participants acknowledged the limitations of reading stand-up text versus hearing it, offering caveats like “*they aren’t amazing on their own but i can imagine how they’d be good in a full stand up routine*” or noting that the “*wording of them is quite strange*”.

Positive Sentiment

Purely positive sentiment was exceptionally rare. When present, it was usually localised to a specific joke or a specific comedic trope that resonated with the individual, such as one participant noting, “*There was quite a bit of self-depreciation in those jokes which was quite funny to read*”, or another praising a specific joke about “*gaining a crippling fear of fluorescent lighting*”.

5.6.8 Semantic Distance

To fully understand the disconnect between the LLMs’ internal optimisation and the external audience reception, the quantitative survey results were directly cross-referenced with the Semantic Distance data from Phase 3.

In Phase 3, the evaluation established that the ‘feedback loop’ successfully acted as a computational editor, pushing the routines away from the extreme “Hallucination” zone (> 0.75) and concentrating them into the theoretical “Sweet Spot” of comedic incongruity ($0.50 - 0.75$) by Iteration 3. If this mathematical theory perfectly mapped to human humour, the audience survey should have heavily favoured the jokes residing in that “Sweet Spot”. However, extracting the exact semantic distance for each of the 260 winning jokes selected by the survey participants reveals something different.

The data indicates that the human participants did actively favour the jokes with the highest semantic distances. The average semantic distance of all the jokes chosen as the ‘best’ or ‘funniest’ by the audience was 0.7693.

When mapped against the established theoretical zones, the distribution of the winning jokes partially aligns with the zones of focus from Phase 3 as shown in Table 20. The main observation that the majority of the winning jokes fall into the “Sweet Spot” and “Hallucination” zones.

Semantic Zone	Winning Votes	Proportion
“Too Predictable” Zone (< 0.35)	0	0.0%
“Light Observation” Zone ($0.35 - 0.50$)	3	1.1%
“Sweet Spot” ($0.50 - 0.75$)	94	36.2%
“Hallucination” Zone (> 0.75)	163	62.7%
Total	260	100%

Table 20: The distribution of audience-preferred jokes mapped against their Semantic Distance zones.

The results presented in Table 20 provide additional context to the votes of the participants, providing a missing link in evaluating computational creativity within this process. A dominant 62.5% of the jokes chosen by human participants were mathematically classified as “Hallucinations” (vectors sharing very little underlying logical similarity). Conversely, only 36.3% of the preferred jokes fell into the targeted “Sweet Spot”.

Semantic Zone	Iteration 1	Iteration 2	Iteration 3	Total Votes
< 0.35	0	0	0	0
$0.35 - 0.50$	0	0	3	3
$0.50 - 0.75$	17	49	28	94
> 0.75	81	55	27	163
Total	97	104	58	260

Table 21: A cross-referenced breakdown of winning audience votes, categorized by the joke’s underlying Semantic Zone and its corresponding iterative stage.

This data perfectly contextualises the earlier finding that audiences preferred the rawer outputs of Iteration 1 (37.7%) over the refined outputs of Iteration 3 (25.6%). In Iteration 1, the LLMs frequently generated highly divergent, absurd, or structurally loose concepts (resulting in high semantic distances). This is shown in Table 21 where the highest vote count in the “Hallucination Zone” was for Iteration 1. During the feedback loop, the LLMs essentially ‘over-correcte” these concepts. Because LLMs operate via probabilistic next-token prediction, the directive to ‘review and improve’ forced the models to mathematically link the setups and punchlines more closely, dropping the semantic distance into the 0.50 to 0.75 range. This is why the vote count for Iteration 2 and Iteration 3 were considerably higher in the “Sweet Spot” as shown in Table 21.

However, by resolving this mathematical distance to satisfy the algorithm, the LLMs partially stripped away the actual human-perceived absurdity. What the vector embeddings measured as a “hallucination” (a distance > 0.75), the human audience frequently perceived as an unexpected punchline, and caused humour. Conversely, what the LLM mathematically registered as a perfect “Sweet Spot” connection, the human participants flagged in their qualitative feedback as *cliché predictable*, or a *word salad* of forced logical leaps.

This cross-referencing provides empirical proof that while Large Language Models are highly capable of manipulating text to achieve a targeted mathematical progression (moving from > 0.75 to a 0.65 distance), this computational proxy for

incongruity does not cleanly map to the subjective human experience of stand-up comedy.

5.6.9 Joke-by-Joke Analysis

To understand the qualitative results and the decisions behind the audience preferences, a joke-by-joke analysis was conducted across the ten joke trios. While the macro-level data indicated a general decline in quality across iterations, examining individual trio performance revealed significant variances based on the thematic content and the specific model used for each joke.

Relatability

The survey results demonstrated a clear audience preference for relatable, observational comedy centered around mundane frustrations and modern domestic life, often referred in the jokes to as “adulting”. This also aligns with the topics being picked in Phase 1 of the generation process. The single highest-rated joke across the entire survey was generated by Google Gemini 2 (Trio 5, Iteration 2), which secured 16 out of 26 votes (61.5%). This joke leveraged the incongruity between adult responsibilities and minor domestic joys: `"Who here gets more excited about their mailbox than their actual significant other? Don't worry, I won't judge... unless your significant other is a golden retriever, then we need to talk"`.

Similarly, Mistral Large 2 achieved a high consensus score (15 out of 26 votes) for a routine anthropomorphising traffic lights, tapping into a universal driving frustration. Claude 3 Sonnet also found success with observational humour regarding the difficulty of assembling IKEA furniture (13 out of 26 votes). The success of these specific topics suggests that Large Language Models possess a strong latent capability for generating observational humour, likely due to their

training data being heavily compiled of internet discourse, memes, and social media posts complaining about these exact ubiquitous experiences.

Struggles with Social Interactions

Conversely, routines focusing on nuanced human interactions, such as **online dating** and **virtual meetings**, showed more volatile results. While initial drafts (Iteration 1) for these topics occasionally performed well, such as GPT-3.5’s opening joke about first dates resembling a **teaser trailer** (15 votes), subsequent iterations for these relational topics degraded rapidly. For example, GPT-4o’s attempt to refine a joke about Zoom calls and networking became overly contrived, and the iterations within the online dating trio split the audience almost evenly, failing to achieve any strong consensus. This indicates that while LLMs can identify the structural framework of observational comedy, they struggle to iteratively refine the subtle, empathetic components required for humour concerning complex social dynamics.

The Trajectory of Iterations

The data revealed two distinct trajectories across the prompt chain. In half of the trios (Trios 1, 2, 3, 7, and 9), the initial generation (Iteration 1) was the undisputed favourite. In the other half (Trios 4, 5, 6, 8, and 10), the model successfully improved the material on its first revision, leading to Iteration 2 securing the most votes. However, as seen with the other qualitative results, Iteration 3 rarely won a majority in the votes, consistently scoring the lowest (e.g., receiving only 3 out of 26 votes in Trios 3 and 7).

A summary of the top-performing individual jokes, highlighting those that achieved the strongest audience consensus, is presented in Table 22.

Model	Topic	Winning Snippet	Votes	%
Google Gemini 2	Adulthood	"Who here gets more excited about their mailbox than their actual significant other?..." (Iteration 2)	16 / 26	61.5%
GPT-3.5	First Dates	"First dates are like a teaser trailer-super glossy, but the actual product..." (Iteration 1)	15 / 25	60.0%
Mistral Large 2	Traffic Lights	"Traffic lights. I swear, they're sentient. They see me coming, and they're like..." (Iteration 1)	15 / 26	57.7%
Claude 3 Sonnet	IKEA Furniture	"It's like trying to decode the Da Vinci Code... You end up with a wobbly coffee table..." (Iteration 2)	13 / 26	50.0%
Mistral Large 2	Driving	"Oh, you're changing lanes? I had no idea! I thought you were just practicing your interpretive dance..." (Iteration 2)	13 / 26	50.0%

Table 22: The top-performing individual jokes across the generated trios in the external survey

Table 23: The winning iterations across all ten generated trios in the external survey, ranked by audience preference.

Model	Topic	Winning Snippet	Votes	%
Google Gemini 2	Adulting	"Who here gets more excited about their mailbox than their actual significant other?..." (Iteration 2)	16 / 26	61.5%
GPT-3.5	First Dates	"First dates... You ever show up and realize they look nothing like their pictures?..." (Iteration 1)	15 / 26	57.7%
Mistral Large 2	Traffic Lights	"Why is it that when you're in a hurry, every single traffic light turns red?..." (Iteration 1)	15 / 26	57.7%

Continued on next page...

Table 23 – Continued from previous page

Model	Topic	Winning Snippet	Votes	%
GPT-4o	Zoom Calls	"Every Zoom call ends with that one person who can't find the Leave Meeting button..." (Iteration 1)	13 / 26	50.0%
GPT-3.5	Wi-Fi / Work	"Why isn't the report done? Uh, well, my Wi-Fi decided to go on vacation..." (Iteration 2)	13 / 26	50.0%
Google Gemini 2	Adulting (Chores)	"Adulting is basically that childhood dream of having your own apartment, except..." (Iteration 2)	13 / 26	50.0%

Continued on next page...

Table 23 – Continued from previous page

Model	Topic	Winning Snippet	Votes	%
Mistral Large 2	Driving (Signals)	"People who don't use their turn signals... Oh, you're changing lanes?... " (Iteration 2)	13 / 26	50.0%
Claude 3 Sonnet	Bank / Taxes	"Every month it's like playing a twisted game of Russian roulette with my bank account..." (Iteration 1)	13 / 26	50.0%
Claude 3 Sonnet	IKEA Furniture	"It's like trying to decode the Da Vinci Code... You end up with a wobbly coffee table..." (Iteration 2)	13 / 26	50.0%

Continued on next page...

Table 23 – Continued from previous page

Model	Topic	Winning Snippet	Votes	%
GPT-4o	Dating Pictures	"Then there are the profile pictures. How many filters can one person use?..." (Iteration 1)	12 / 26	46.2%

5.6.10 Shortcomings

This experiment both supports and contrasts the findings of the previous internal metrics. While the models successfully followed the mechanical instructions to ‘review’ and ‘revise’, the data suggests several underlying reasons why this computational editing process ultimately degraded the comedic quality for human readers.

The “Thrashing” and Over-Editing

As identified during the evaluation of the Levenshtein Distance, the LLMs exhibited a significant inability to execute minor, surgical edits. In human comedy writing, late-stage refinement focuses on tweaking rhythm, syllables, and pacing. In contrast, the AI maintained an exceptionally high average edit rate of 43.42% between Iteration 2 and Iteration 3.

Lacking the contextual memory to isolate and fix only the flawed elements, the models assumed the entire document required rewriting. This “thrashing” effect directly correlates with the “Chaos” noted by survey participants. By continuously tearing down and rebuilding the routine’s foundation, the LLM destroyed the

natural pacing and narrative flow established in the initial drafts.

Regression to the Mean

LLMs operate via autoregressive next-token prediction, meaning they are inherently optimized to generate the most plausible and expected continuations of text. During the initial “engagement” phase (Iteration 1), the models may generate novel, surprising, or highly specific text based on the initial prompt. However, as the model is repeatedly asked to ‘review’ and ‘fix’ its own work, it may trend towards a more ‘average’ result, as it sees the ‘correct’ response as the response that is most common within its dataset.

Faced with a directive to correct errors, the probabilistic engine defaults back to the safest, most widely represented patterns in its training data. This explains the “Cliché” issues identified in the qualitative feedback. The iterative ‘feedback loop’ essentially smoothed out the unique edges of the jokes, forcing the output back toward recognizable, well-worn tropes that the LLM mathematically recognised as ‘correct’ comedy, but which humans perceived as derivative or boring.

“Mathematical Incongruity”

In Phase 3, the Semantic Distance evaluation proved that the models successfully steered their text into the mathematical “Sweet Spot” (0.50 to 0.75 distance) by Iteration 3. However, the audience survey suggests that mathematical semantic distance does not perfectly translate to human comedic incongruity.

To achieve this required vector distance, the models often forced logical leaps that made sense mathematically but failed semantically for a human reader. For example, connecting `Wi-Fi issues` to `summoning a genie` might satisfy the AI’s algorithm for surprise and distinctiveness, but to a human audience, it reads as a nonsensical non-sequitur (as noted by Participant 3). The models optimised for a mathematical proxy of humor, missing the vital cultural and contextual glue

required to make the incongruity actually land.

The “Text-Performance” Gap

Stand-up comedy is an embodied, live art form. When human comedians edit their work, they do so by visualising the live delivery, in a performance context. The LLM, however, edits purely as a literary engine. It adds complexity, multiple punchlines, and structural twists that look impressive on paper as a block of text, but read as overly wordy and conceptually disjointed when a human attempts to process them as spoken dialogue. This may account for some of the jokes performing badly, as they weren’t being performed by a comedian on stage. This idea was noted by Participant 17.

5.6.11 Statistical Significance

While the percentage distribution of the aggregate votes indicates a preference for earlier iterations, it is necessary to determine if this variance is statistically enough to draw definitive conclusions about human comedic appreciation. To test this, a Chi-Square Goodness of Fit (χ^2) test was conducted on the Wave 1 results.

The null hypothesis (H_0) suggests that there is no true audience preference among the three iterations; therefore, the votes should be distributed equally across all three options (approximately 33.33% per iteration).

Based on the 26 participants completing the mandatory 10 joke trios in Wave 1, the total sample size was 260 votes ($N = 260$). Applying the aggregate percentages to this total yields the observed raw frequencies, which were compared against the expected frequencies under the null hypothesis (86.67 votes per iteration). The results of the Chi-Square calculation are detailed in Table 24.

The calculation yielded a total Chi-Square value of $\chi^2 = 6.74$. With 3 categories, there are 2 degrees of freedom ($df = 2$).

To ensure rigorous analysis and prevent Type I errors (false positives) in a

Stage	Observed (O)	Expected (E)	%	χ^2
Iteration 1	98	86.67	37.7%	1.48
Iteration 2	95	86.67	36.7%	0.80
Iteration 3	67	86.67	25.6%	4.46
Total	260	260.00	100%	$\chi^2 = 6.74$

Table 24: Chi-Square Goodness of Fit test results comparing observed audience preferences against an expected uniform distribution ($N = 260$).

highly subjective domain like comedy evaluation, a stringent alpha level of $\alpha = 0.01$ (a 99% confidence interval) was applied to the results. At this threshold, the critical value required to reject the null hypothesis is 9.21.

Because our calculated value of 6.74 is significantly lower than the critical value of 9.21 (resulting in a p -value of approximately 0.034), the null hypothesis cannot be rejected at the 99% confidence level.

It suggests that while Iteration 3 received fewer overall votes, the variance across the three iterations is not robust enough to claim a definitive, universal rejection of the LLM’s refinement process. The audience’s preferences, while leaning slightly toward the unrefined drafts, lacks the statistical magnitude required to prove that the automated editorial loop consistently ruins the subjective quality of the jokes. This aligns with the findings from Wave 2 and the per-model breakdowns, which demonstrate that Iteration 3 can and does perform competitively.

5.6.12 Summary

Ultimately, the external audience survey serves as a counterweight to the internal computational metrics evaluated earlier in this chapter. The findings present a clear paradox in computational creativity. While Phase 3 demonstrated that large language models are highly capable of executing the mechanics of targeted self-improvement, this computational self-reflection does not equate to genuine comedic improvement.

The quantitative rejection of Iteration 3 by the survey participants, coupled with the qualitative feedback regarding disjointed narratives, cultural disconnects, and semantic confusion, highlights the current limitations of LLMs in highly subjective creative domains. The models effectively treated the refinement process as a mathematical optimisation task. The models aggressively rewrote text to action specific feedback and successfully pushed semantic distances toward a theoretical “sweet spot”. However, they stripped away the natural pacing, cultural relevance, and cohesive narrative flow required to make a human audience laugh.

However, this aggregate perspective ignores the individual success of specific models. The ‘feedback loop’ of some models (and architecture) frequently achieved its intended goal. Models such as GPT-3.5 and GPT-4o demonstrated substantial alignment with human preferences. This proves that under the write conditions, specific models can exhibit “Progression and Development” to create better stand-up comedy material. This is supported by the significance test, which showed that the preference away from Iteration 3 was not statistically significant for the highly subjective domain of dealing with humour and individual comedic preference.

This external validation leads to a nuanced conclusion regarding the generative capabilities of LLMs within the domain of stand-up comedy. While the models exhibit a clear, measurable capacity for the “Progression and Development” (Jordanous and Keller 2016) of a text from a strictly computational standpoint, the artistic output is often heavily degraded when subjected to a rigid, automated, multi-stage editorial loop. The LLM can successfully act as its own editor mechanically, but it potentially lacks the contextual and emotional awareness to recognise when those mechanical edits have ruined the joke. Nevertheless, these results demonstrate that bridging the gap between computational editing and human humour is possible, even if the current methodology yields inconsistent results.

5.7 Chapter Summary

Throughout all steps in this simulated creative pipeline, the core objective was to determine whether Large Language Models can demonstrate true creativity when generating content in the highly subjective, structurally demanding domain of stand-up comedy. To identify this, the “Progression and Development” (Jordanous and Keller 2016) metric was utilised as it directly parallels the way human performers use iterative writing, critical review, and performance tagging to produce their best work. The four stages of this process were individually analysed for evidence of these creative skills.

5.7.1 The Engagement-Reflection Model

The success of this progression can be mapped directly to the Engagement-Reflection model of creative writing (Pérez y Pérez 1999). During the initial “engagement” phase (Phase 2: Material Generation), the system freely generates candidate material driven by constraints but without explicit critical judgment. The transition into the “reflection” phase (Phase 3: The ‘Simulated Feedback Loop’) is where the LLMs truly demonstrated their capacity for computational creativity, paralleling the human process of refining and ‘workshopping’ material.

5.7.2 Overview of the experiment

When observing the dataset as a whole, the aggregated generations provide strong mathematical proof of this progression. Initially, 13 of the 25 generations fell into the nonsensical “Hallucination” zone. However, by the final iteration of the feedback loop, the average of the generations shifted dramatically, with 17 of the 25 routines successfully landing in the optimal “Sweet Spot” of comedic incongruity. Across the board, four out of the five tested models achieved an average semantic distance within this ideal target zone. Furthermore, when crossing the text-performance

gap, 88% of all generations successfully demonstrated progression by applying entirely unique performance annotations not found in the prompt’s sample list.

The data reveals that different models exhibit drastically different strengths across the creative workflow, effectively averaging into a comprehensive, though varied, display of artificial creativity:

1. **Material Generation:** All models successfully progressed beyond the basic prompt, implementing unprompted structural comedy mechanics (Callbacks, Rule of Three, Superiority Formula). They proved capable of replicating the syntax of comedy and successfully establishing “Script Opposition” (Attardo and Raskin 1991).
2. **Self-Review and Improvement:** The models proved highly capable of acting as their own ‘internal auditors’. Despite exhibiting a tendency toward aggressive over-editing (high Levenshtein distances) rather than surgical polishing, the models demonstrated high intentionality. They successfully identified structural weaknesses (e.g., weak punchlines, lack of audience interaction) and directly actioned those critiques in subsequent drafts, frequently pushing their text into an optimal “semantic sweet spot” of incongruity.
3. **Performance Annotation:** Finally, the models successfully bridged the text-performance gap. While Mistral 2 Large proved to be the most prolific annotator (highest volume), it relied heavily on safe, provided examples. Conversely, Google Gemini 2 was the most inventive, generating the highest percentage of unique, unprompted stage directions. Ultimately, GPT-3.5 and GPT-4o found the most effective middle ground, producing the most qualitatively detailed and structurally consistent performance scripts.

External Validation

The external validation conducted via an audience survey presented an alternative perspective to these internal computational metrics. While the models successfully executed the mechanics of self-improvement, human participants sometimes rejected the final polished routines, in favour of earlier drafts. The survey revealed an inverse relationship between iterative refinement and audience preference, with participants favoring the rawer outputs of Iteration 1 (37.7%) and Iteration 2 (36.7%) over the targeted Iteration 3 (25.6%). Qualitative feedback indicated that the models’ aggressive editing loops often led to reduced structural coherence, replacing natural pacing with disjointed narratives, over-localized clichés, and logical leaps that confused rather than amused. This disconnect highlights a limitation: while LLMs can computationally mimic the “progression” of creative editing, they currently lack the contextual awareness necessary to gauge when a mechanical revision compromises the actual human experience of humor.

However, a more detailed per-model breakdown, shows that this aggregate data obscures significant successes within specific models and architectures. Some models demonstrated substantial alignment with human preference, outperforming the earlier baseline and refinement iterations.

In summary, the models did not simply engage in next-token prediction; they exhibited intentionality, contextual memory, and a demonstrable ability to reflect upon and improve their own creative outputs over time. The success of this multi-phase process highlights that the increased complexity of modern Large Language Models, when guided by targeted, iterative prompting, allows them to effectively model genuine creative generation within the nuanced domain of stand-up comedy. Furthermore, the isolated successes observed in the individual model data prove that the automated editorial loop is “developing” the routines and is not universally detrimental. Rather, it highlights that bridging the gap

between computational editing and human humour is entirely possible, providing evidence that iterative self-reflection can indeed demonstrate Creativity through the “Progression and Development” component of creativity.

Chapter 6

Discussion

6.1 Successes of the Research

It introduced a prompt engineering methodology that was used to guide several public closed-source large language models through a process of generating, critiquing, and improving stand-up comedy material. This methodology was used to generate a selection of results that could be analysed and discussed. Beyond this, the experiments were designed to identify evidence of “progression and development”, a key component of creativity in computational systems (Jordanous and Keller 2016).

Interesting and unseen aspects of this LLM technology were also identified, such as:

- The system displaying clear signs of intentionality through the improvements it makes.
- An ‘understanding’ of stand-up comedy as a genre through the techniques it uses and identified.
- An understanding of ‘bad’ and ‘good’ through the experimentation with creating a worse stand-up comedy routine, through the iterations.

As shown through the Evaluation chapter of this thesis, the models’ performance varied from model to model, but all the models showed evidence of this “progression and development” ability throughout the experiment. The prompt-engineered process resulted in strong responses from the LLMs and could be adapted in the future to be used on other written-based media. The results obtained provided clear and repeatable evidence that could be investigated using different domains, other than stand-up comedy.

6.1.1 Demonstrating “Progression and Development”

This thesis aimed to investigate the self-improvement capabilities of large language models through the domain of stand-up comedy. The primary research question asked whether LLMs, using prompt-chaining workflows, could successfully model self-review and self-improvement. The results from the evaluated dataset provide sufficient evidence of this. Through the structured prompt-chaining methodology, the models successfully navigated the process of generating, reviewing and improving their own stand-up comedy material. In doing so, the models provided clear, repeatable evidence of “Progression and Development” which serves as a vital empirical component for attributing creativity to computational systems (Jordanous and Keller 2016).

The success of this iterative refinement is directly tied to the cognitive framework of creative writing known as the “Engagement-Reflection” cycle (Pérez y Pérez 1999). LLMs operate continuously in the “engagement” phase, generating material freely by predicting the next most statistically probable token. When not guided by explicit prompting, this continuous engagement results in stochastic, average text that (in the domain of stand-up comedy) reads like ‘open-mic level’ comedy, as the model defaults to the statistical mean of its training data without critical judgment.

However, the automated ‘feedback loop’ introduced in the methodology successfully forced the LLMs out of this continuous generation loop. By chaining prompts to demand self-critique and scoring, the models were forced into the “reflection” state. This forced pause allowed the models to act as their own ‘internal auditors’. It enabled them to identify structural flaws, tighten their own punchlines, break creative impasses, and purposefully action specific feedback in their subsequent drafts.

Looking through the lens of psychological versus historical novelty shows the further success of the models within this domain. Due to LLMs being probabilistic engines trained on vast datasets of existing human knowledge, it is computationally intractable for them to achieve “H-Creativity”. They cannot truly invent a human experience that has never existed before in the history of the domain (Boden 2004). This research demonstrates that they do not need to achieve H-Creativity to be deemed creative within the bounds of a session. By taking a starting premise and successfully evolving it across multiple iterations (i.e. by shedding clichés, subverting expectations, and tightening narrative arcs) the models proved they can achieve “P-Creativity” (psychological novelty). The text the models produced at the end of the refinement loop was novel relative to the model’s own starting point, satisfying the criteria for genuine computational creativity through progressive, internal development.

6.1.2 Success through Semantic and Structural Validity

While the prompt-chaining methodology proved the models were capable of engaging in an iterative process, evaluating the actual text generated requires a deeper look at the structural mechanics of the jokes. An essential finding of this research is that the LLMs successfully manifested “Script Opposition” without explicit structural prompting, thereby proving a high level of domain competence. According to the General Theory of Verbal Humour (GTVH), humour cannot occur

without the presence of two overlapping but opposing scripts (Attardo and Raskin 1991). Throughout the generations, the models successfully mapped mundane, highly relatable experiences onto conceptually distant scenarios (such as equating a **self-checkout machine** to the **Bermuda Triangle**, or **dating apps** to **Amazon shopping**). Because the prompt provided no specific examples of stand-up techniques, the AI’s ability to natively establish these oppositions demonstrates an inherent understanding of comedic structure.

This success is best understood through Tony Veale’s concept of “Linguistic Readymades” (Veale 2013). Veale argues that computational wit does not require deep human consciousness, but rather the ability to retrieve and recontextualise existing cultural artifacts (Veale 2012). In this study, the LLMs acted precisely as curators of these readymades (Veale 2021). The models harvested universal tropes from their training data (such as the **frustrations of customer service**, the **quirks of family members**, or the **awkwardness of online dating**) and mashed them together to create incongruity.

However, measuring this incongruity mathematically revealed the models’ initial struggle with balancing surprise and logic. As mapped through cosine similarity and semantic distance, early drafts frequently fell into either the “Too Predictable Zone” (< 0.35), where jokes read as factual observations, or the nonsensical “Hallucination Zone” (> 0.75), where the semantic gap was too wide to offer a logical resolution. The reflection phase of the methodology acted as a computational editor that reined the models in. The data proved that as the models progressed from Iteration 1 to Iteration 3, they actively adjusted their semantic choices to balance high distinctiveness (the surprise) with a mathematically sound connection (the resolution). By the end of the refinement loop, the majority of the routines (17 out of 25) were successfully guided out of hallucination and anchored into the optimal comedic “Sweet Spot” (0.50 to 0.75).

In one of the experiments while engineering the prompts, the models were

tasked with deliberately making a routine “worse”. If a model understands what makes a joke work, it must also understand how to break it. The LLMs successfully interpreted this instruction by systematically dismantling the GTVH parameters: they removed punchlines, inserted awkward conversational fillers (`let’s talk about... um... trees`), and explicitly encoded stage directions of audience silence and abandonment. This ability to intentionally sabotage a routine provides evidence of a deep, structural comprehension of failure within the stand-up comedy domain. This proves that the adjustments made to the stand-up routines across iterations were driven by a semantic understanding of comedy mechanics rather than stochastic noise.

6.1.3 Prompt Engineering

The development and application of the prompt engineering methodology was a key aspect of this research, revealing insights about how users must interact with LLMs to produce genuinely creative outputs. The decision to bring together specific elements aligning with the SMART and CLEAR frameworks, alongside ‘persona prompting’ was essential. By combining explicit, actionable instructions with a ‘simulated self’, the prompts acted as a cognitive scaffold (similar to the scaffolding provided to pupils when learning new concepts in school). The persona targeted the model within the goal domain (stand-up comedy) while the instructions ensured the outputs were organised for analysis.

In a zero-shot environment, without a persona, the models drifted towards the statistical mean of all text resulting in dry, explanatory and highly literal outputs. The engineered prompts narrowed the outputs and increased the probability of generating comedic structures (such as setups, punchlines and incongruity). The prompt engineering was the mechanism that transformed the model from a general-purpose text predictor into a domain-specific agent

A highly successful aspect of the prompt design was how formatting instructions

inadvertently triggered complex evaluative behaviors. For example, explicitly requiring the models to output their reasoning under a specific subheading titled *Justification* did more than just organize the text for analysis. It actively forced the model to generate the reasoning behind its choices. This structural prompt command effectively activated the “reflection” phase of the Engagement-Reflection cycle (Pérez y Pérez 1999). It demonstrated that prompt engineering for computational creativity is not just about telling the model what to write, but planning how the model evaluates its own outputs.

Upon reflection, this research also showed how users need to respond to the models’ emergent behaviours, for example, randomly inserting state directions into the routines. These unpredictable inconsistencies of LLMs led to specific decisions in the workflow created during the methodology, where these paralinguistic features could be intentionally encoded into the routines during Phase 4. This reflects the real-world practicalities of working with generative AI. Prompt engineering is not only the preliminary writing and setup phase, but (to be successful) must be a continuous responsive dialogue between the users’s intent and the model’s outputs.

6.2 Limitations of the Research

Although successfully showing evidence to support that LLMs can self-review and self-improve themselves to produce better outputs according to some metrics, such as Semantic Distance, qualitative use of stand-up techniques and the Levenshtein Edit distance, it is important to reflect on the limitations of this prompt-chaining approach. A central challenge throughout these experiments was the tension between granting the model creative freedom (using zero-shot or few-shot prompts) versus restricting it to ensure consistent, measurable data.

6.2.1 Consistency

In order to give the LLMs the most ‘freedom’ to organically explore and evolve their responses, the prompts used in the main study were kept relatively broad, and limited in terms of samples or examples. Adding examples to a prompt can result in the LLM drawing from the example and, in some cases, directly plagiarising. Although, on the other hand, it helps provide a frame of reference for the LLM to use, often leading to better results. Balancing this proved an interesting challenge in this thesis, as there needed to be enough context in the prompts for the LLMs to use, but not enough that it plagiarises or copies.

Sample content was only used during Phase 4 of the process (adding annotations). These annotations were still generated by the LLMs but were passed to each LLM within the prompt and led to a majority of the annotations being copied from the list, rather than being “developed”. In all other instances of review criteria (i.e. what to look for in a stand-up comedy routine), the LLMs were allowed to decide the criteria themselves. Although showing more evidence of intentionality and ‘understanding’ in the answers, it means that there is a distinct lack of a control variable that is consistent between routines to use in the results.

An alternative experiment demonstrated that heavily constraining the prompt yields far more consistent, but less creative, results. For example, an alternative prompt that would have provided more consistent reviews across the models, forcing it to explicitly check for specific techniques, would be:

Prompt 6.1: An alternative prompt for reviewing the stand-up comedy routines using a more consistent approach

I would like you to review this stand-up comedy routine ,
 taking into account this list of criteria: – Has the
 routine stuck to the required topic? – Has it applied
 techniques such as ‘Callbacks?’, ‘The Superiority
 Formula’, ‘Incongruity’ and ‘The Rule of Three’ – Have
 all the setups had effective punchlines? – Has the
 routine made use of tags after the punchlines? Identify
 the advantages of the routine and suggest improvements
 that would make the routine better. After this , I would
 then like you to rate the stand-up routine out of 10,
 using the criteria . Having done this , in a subheading
 called ‘Justification ’, tell me how you have calculated
 this score , using the criteria .

After applying this highly rigid evaluation prompt to the routine in generation GxDx01, using the GPT-4o model, the resulting review shifted dramatically from organic critique to a highly structured, robotic assessment. Rather than generating a fluid, conversational review, the model produced a rigid breakdown that strictly adhered to the constraints of the prompt.

An analysis of the model’s output reveals exactly how these strict parameters altered its behaviour:

- The model successfully parsed the routine for the exact comedic theories requested. It accurately identified incongruity (highlighting a joke that compared dating apps to **Amazon shopping**) and successfully isolated a Rule of Three structure within the text ("Swipe left, swipe right... Two-day shipping?").

- Because the prompt explicitly demanded a check for callbacks, the model actively noted their absence. Rather than glossing over the missing technique, the LLM used this absence as a primary point of critique, suggesting that adding a callback "would strengthen the routine's cohesiveness and provide a satisfying payoff."
- Following the instruction to rate the routine out of 10, the model awarded an 8/10 and generated a strict sub-rubric to justify the score. It actively broke down the available points based on the prompt's checklist, scoring **Topic Adherence** a 2/2, but dropping **Techniques** to a 3/4 specifically because of the missing callback.

While this alternative prompt successfully solved the issue of consistency (providing a highly uniform, easily measurable review output across generations) it did so at the cost of organic creativity. It severely limited the model's organic input. By forcing the LLM to act as a strict examiner marking a rubric, the prompt removed the opportunity for the AI to use its own definitions of 'good' stand-up and restricted its ability to progress and develop the routine beyond the strict, user-provided boundaries.

Consistency could be applied to other areas of the experiments as well, namely in the routines being generated. In these experiments, the model generated from start to finish each time, choosing a topic and then building a routine from that topic. This resulted in vastly different routines from each generation, varying in topics, length and style. An additional experiment could have been to develop the same routine with different models to compare the changes each model made and how different each model's routine ended up.

6.2.2 The “Ship of Theseus” Problem

Another significant limitation relates to how the models executed their rewrites. When human comedians refine a routine, their workflow typically moves from broad, structural rewrites (an ‘axe’) in early drafts, down to surgical, word-level adjustments (a ‘scalpel’) during the final polish. However, the longitudinal data (measured via the Normalised Levenshtein Distance) demonstrates that the LLMs fundamentally struggled to ‘put down the axe’.

In 64% of the experiments, the models exhibited an instability or “thrashing” pattern. Lacking the contextual memory to execute surgical edits, the models frequently assumed the entire document was flawed and aggressively rewrote it, resulting in routines that changed by 90% or more across the three iterations. This presents a philosophical and practical limitation akin to the “Ship of Theseus” thought experiment. If an LLM replaces almost all the words in a routine, is it actually refining a draft, or just generating a brand new one? Consequently, it brings into question whether rewriting 90% of a text constitutes true intentionality, a requirement for creativity according to Ventura (Ventura 2016), or if the model is simply defaulting back to mere generation by writing a completely new sequence of tokens.

The audience survey confirmed the negative artistic impact of this computational “thrashing.” In qualitative feedback, participants highlighted a theme of **structural decline** and **chaos** in the later iterations. Because the LLMs rewrote and rebuilt the routines rather than making surgical edits, the final drafts frequently abandoned premises prematurely and lost the natural pacing established in Iteration 1. By constantly replacing the ‘planks’ of the routine, the LLM destroyed the narrative flow that human readers rely on to track a joke.

6.2.3 Limitations of Computational Metrics

While the triangulation strategy provided empirical grounding, computational metrics have inherent limitations when applied to subjective art forms. Semantic Distance (Cosine Similarity) serves as an excellent proxy for measuring incongruity and the surprise factor of a joke, but it cannot fully capture the subjective human experience of what is actually ‘funny’.

This limitation was shown through the external audience survey. While the Semantic Distance metrics proved that the LLMs successfully guided their jokes into an optimal “Sweet Spot” (between 0.50 and 0.75) by Iteration 3, human participants actively rejected these iterated drafts. The survey data showed an inverse relationship between computational refinement and audience preference, with only 25.6% of participants voting for the final Iteration 3 routines, compared to 37.7% for the raw Iteration 1 baseline.

What the algorithm perceived as a successful, mathematically sound semantic leap (e.g., connecting a `Wi-Fi outage` to `summoning a genie`), human readers frequently perceived as a confusing non-sequitur or a loss of narrative coherence. The models successfully optimized for a mathematical proxy of humour, but in doing so, they stripped away the natural pacing and cultural grounding required to make the incongruity land with a real audience. This proves that while computational metrics are vital for tracking system behaviour and text manipulation, they cannot currently serve as a standalone barometer for comedic quality.

Evaluating the routines solely through text overlooks the “Text-Performance Gap” identified by Schwarz (Schwarz 2010). A stand-up routine is an incomplete artifact until it is performed live. While Phase 4 of the methodology successfully proved that LLMs can encode paralinguistic markers (such as stage directions, tone, and physical act-outs), they cannot truly execute them. Elements such as intonation, the precise timing of the “comic pause,” and dynamic rhythm adjustments are inherently bound to real-time human delivery. An LLM can

script a pause, but it cannot adjust its rhythm to account for the actual “laughter decay” of a live audience. Therefore, evaluating the abilities of these models via written annotations remains a theoretical measurement of potential, rather than a definitive assessment of stage-readiness.

6.2.4 Qualitative Analysis

While this research utilised qualitative analysis to identify evidence of intentionality and genre understanding, the use of such a subjective lens in the domain of stand-up comedy introduces some methodological friction. Unlike quantitative metrics (such as Cosine Similarity) qualitative evaluation is tied to the researcher’s personal ‘humour filter’.

Comedy is perhaps the most subjective art form; what one evaluator identifies as a “subversion of expectation,” another may dismiss as a “hallucinatory non-sequitur.” By relying on qualitative interpretation to determine if a routine improved, there is a risk of researcher bias. In an effort to validate the prompt-engineering methodology, the researcher may subconsciously over-index on subtle improvements while overlooking persistent structural failures, effectively finding the progression they set out to prove.

A robust qualitative study would typically requires multiple “judges” to reach a consensus (Inter-Rater Reliability). Because this study’s qualitative insights were primarily driven by a single point of view, the findings lack the “stress test” of a diverse audience. A joke about `dating apps as Amazon shopping` might be qualitatively coded as “structurally sound” by a millennial researcher, but could be coded as “incoherent” or “boring” by a different demographic.

Qualitative analysis allows for a deep, surgical look at specific routines (e.g., `GxDx01`), but it does not scale. While this research identified evidence of “P-Creativity” in a select dataset, qualitative methods cannot definitively state whether these successes are repeatable across the billions of potential prompt

permutations. It trades breadth for depth, leaving open the question of whether the successes identified are representative of the model's core capabilities or are merely 'cherry-picked' outliers in a collection of average generations.

Further analysis using a wider sample of reviewers and more detailed quantitative evaluation methods, would tackle many of these considerations, but falls outside the scope of this project.

6.2.5 Limitations of the Survey

A small limitation of the external survey methodology was the reliance on extracting and evaluating individual jokes in isolation from their wider generated routines. Stand-up comedy is inherently highly contextual. Its success depends heavily on narrative momentum, the establishment of a consistent comedic persona, and structural devices such as delayed callbacks that span several minutes of material. By isolating a single setup and punchline, the survey removes the joke from its structural ecosystem. Consequently, participants were judging the micro-level semantic humour of a fragment rather than the macro-level coherence and pacing of the full routine. This removal of larger context slightly limits the participants' ability to judge the effectiveness of the jokes in the context of the routines. Qualitative feedback from the survey acknowledged this deliberate boundary, with participants noting the necessity of human delivery and broader context to elevate the material. For instance, respondents observed that the jokes "aren't amazing on their own but i can imagine how they'd be good in a full stand up routine", and that "as is typical of AI, they're mostly not funny. Imagining them performed by a human using appropriate tone of voice, body language, and gestures makes some of them funnier". Ultimately, as another respondent summarised, "they might be better when delivered properly in a stand-up setting rather than

reading, but as it is these jokes are mediocre at best, none stand out as genuinely hilarious".

Furthermore, using a blind A/B/C survey format presents inherent constraints when applied to the performative arts. Stand-up comedy is traditionally a dynamic, multi-modal experience characterised by verbal timing, physical gesturing, and the contagious nature of communal laughter. Reducing this live experience to a static textual comparison forces participants into a highly analytical mode of reading that contradicts the natural, spontaneous reception of humour. Additionally, forced-choice A/B/C testing compels a definitive preference, potentially masking nuanced qualitative reactions where a participant might find all three variations equally lacking or subjectively funny in entirely different ways. This anticipated constraint was reflected in participant feedback, where respondents expressed a desire for a negative option, stating: "They were all a bit similar. Needs to be an option for none of these are funny." Compelling a selection meant participants were sometimes effectively choosing the least flawed iteration rather than a genuinely successful joke, exemplified by the comment: "None of the jokes were really that funny and I just selected the better one out of the three."

Despite these necessary compromises, this evaluation framework was deliberately selected as the most rigorous and effective way to maximise genuine engagement from the general public. In the domain of stand-up comedy, the general audience operates as the ultimate decider of success. In many ways, they are the true domain experts whose spontaneous reactions dictate whether a comedian, or a joke, is actually funny. Gathering statistically meaningful feedback therefore required accessing a broad demographic reflective of an authentic comedy audience. Demanding that volunteer participants read, analyse, and directly compare three complete, multi-page stand-up routines would have imposed an unreasonable cognitive load and time commitment. Such a demand would have inevitably resulted

in severe survey fatigue and high abandonment rates, ultimately compromising the validity of the data. Therefore, extracting individual jokes and presenting them in a rapid, comparative survey format served as a vital methodological trade-off. It sacrificed the long-form and performance-based context in exchange for generating a robust, accessible dataset of authentic audience preferences, that fit within the scope of this research project. It ensured that the models were judged by the very people stand-up is designed to entertain, without overburdening the participants.

6.3 Ethical Implications of AI-generated Comedy

The findings of this study develop existing ethical discussions around AI-generated comedy (as mentioned in Section 3.5 of the literature review) by moving beyond purely theoretical concerns and grounding them through observable system behaviour. While the literature establishes key tensions in these discussions, such as authorship, lived experience and moderation, the evaluation in the previous chapter demonstrate how these issues manifest in practice when large language models (LLMs) attempt to generate, critique and refine stand-up comedy.

Rather than being abstract risks, the results reveal that this ethical questions are concerns around the generative process itself. In particular, the concepts of “hallucinated biography”, constrained violation, presenting these questions as observable limitations that shape the form, quality, and implications of AI-generated humour.

6.3.1 The “Hallucinated Biography” Revisited

The literature frames stand-up comedy as an autobiographical art form, rooted in a performer’s lived experience and social positioning. This study reinforces that claim by demonstrating how LLMs attempt to simulate this autobiographical voice without possessing any underlying self.

Across the resultant dataset, models consistently adopted first-person framing, constructing routines around **dating**, **family**, **work**, or **social anxiety**, despite the models having no life experience for grounding these topics. This aligns with the concept of a “hallucinated biography”, but the ‘feedback loop’ process adds an important nuance. These fabricated (not real) perspectives are not static. Through the process, the models increasingly refine these pseudo-identities, adding specificity, anecdotes, and even personal reflections.

For example, during the refinement phase, models responded to feedback such as **add a personal anecdote** by generating increasingly detailed scenarios, such as **gym experiences** or **awkward social encounters**. This demonstrates a form of simulation, where the system learns to approximate the structure of lived experience rather than its substance or grounding within real-world experience.

However, this creates an ethical tension. While the results show that specificity improves perceived comedic quality (particularly through movement toward the semantic “Sweet Spot”), the authenticity underpinning that specificity remains entirely fabricated. The audience is therefore engaging not in a traditional suspension of disbelief (as they would with a human comedian), but in a knowing acceptance of artificial fabrication.

This raises a critical question for future research: if relatability in stand-up comedy is built on authenticity, does an increasingly convincing simulation constitute a form of deception, or does it redefine authenticity for computational systems?

6.3.2 Censorship, Safety and Violations

The evaluation provides empirical support for concerns raised in the literature regarding moderation and its impact on humour. While the Benign Violation Theory (BVT) suggests that humour requires a balance between norm violation and safety, the results indicate that LLMs struggle to consistently maintain this

balance, particularly under the filtering systems used by these publicly accessible models.

In Phase 1 (Topic Generation), the models overwhelmingly selected ‘safe’ and highly typical domains (81.6% aligned with established comedy categories), such as `technology`, `travel`, and `relationships`. While this demonstrates strong domain knowledge and awareness, it also reveals a narrowing of comedic scope toward culturally neutral and broadly acceptable topics.

This trend persists in the material generation phase. Although the models successfully employed recognised comedic techniques (e.g., rule of three, callbacks, self-deprecation), these were often applied within constrained thematic boundaries. The result is humour that is structurally valid but typically reads as ‘safe’.

The semantic distance analysis further clarifies this limitation. While iterative refinement successfully moved many jokes into the “Sweet Spot” of incongruity (0.50 – 0.75), this optimisation often involved reining in more extreme or abstract ideas rather than pushing boundaries. In other words, the ‘feedback loop’ acts as a form of internal moderation, guiding outputs away from both incoherence and potential risk.

This suggests that LLMs do not simply remove harmful content; they systematically compress the ‘violation’ component of humour into a narrow, acceptable range. The risk here is homogenisation, where the comedy becomes optimised for safety and coherence rather than disruption. This could be why a lot of LLM generated comedy is critiqued as ‘corporate’ or ‘sanitised’.

6.3.3 Creativity, Iteration and Intentionality

This experiment shows clear insight into the iterative refinement process for LLMs. The results show clear evidence of “Progression and Development”, with models demonstrating the ability to critique and improve their own work across multiple iterations.

At first glance, this appears to challenge the notion that LLMs lack intentionality. The models not only identified weaknesses (e.g., **weak closers**, **lack of audience interaction**) but also implemented targeted improvements, such as adding callbacks, tightening punchlines, or introducing act-outs.

However, the quantitative analysis complicates this ethical interpretation. The high Levenshtein edit distances (often exceeding 40–50% between iterations) reveal that many improvements involved substantial rewrites rather than incremental refinement. The previously mentioned “thrashing” pattern suggests that the system is not editing with precision but regenerating content at scale.

The “Ship of Theseus” analogy becomes particularly relevant here. If a routine is rewritten by over 90% across iterations, it raises the question of whether the system is refining an idea or continuously replacing it. From an ethical perspective, this challenges the attribution of creativity: is the model demonstrating reflective authorship, or simply producing successive approximations of plausible outputs?

6.3.4 Cultural Ownership

The results also provide new insight into debates surrounding cultural ownership and originality. The high degree of typicality observed in topic generation (81.6% alignment with established categories) confirms that LLM outputs are deeply rooted in existing comedic conventions.

At the same time, the models demonstrated clear evidence of “P-creativity”, particularly through increased specificity and unexpected combinations of ideas. For example, prompts related to **social media** or **family life** were frequently developed into highly specific and theatrical premises, indicating the system’s ability to recombine familiar elements in novel ways.

However, this novelty operates within a constrained space. The reliance on common themes, combined with the probabilistic nature of generation, results in outputs that feel familiar even when they are technically new. This aligns with the

concept of the model as a ‘stochastic parrot’, but the evaluation provided a closer look at this idea. Rather than simply repeating patterns, the model reconfigures them within statistically likely boundaries.

The ethical concern, therefore, is not direct plagiarism but diffuse appropriation. The comedic voice produced by the model is an aggregate of countless uncredited influences, raising questions about ownership, compensation, and the erosion of individual artistic identity.

This becomes particularly important when considering ‘style transfer’. Although not explicitly tested in this study, the demonstrated ability of models to internalise and reproduce domain-specific techniques suggests that ‘style transfer’ is not only possible but embedded in the generation process.

6.3.5 Summary

Taken together, these findings suggest that the ethical implications of AI-generated comedy go beyond what was discussed in the literature, and cannot be reduced to issues of bias, plagiarism, or censorship alone. Instead, they point to a broader shift in how humour is produced, structured, and understood.

LLMs do not simply imitate stand-up comedy. They build it as a system of probabilistic patterns, optimisation processes, and iterative refinements. In doing so, they expose the underlying mechanics of humour and stand-up comedy theory. They do this, whilst simultaneously stripping away the human context that gives those mechanics meaning.

The central ethical tension when creating stand-up comedy with an LLM, is not whether AI can be funny, but whether humour can retain its cultural and personal significance when detached from lived experience, risk, and accountability. In other words, when detached from ‘a person’.

6.4 Future Work

While this research successfully established a framework for generating and refining stand-up comedy through prompt engineering, several avenues for future exploration remain to fully test the capabilities of Large Language Models in this domain.

This thesis has explored how LLMs can improve themselves to produce better material, emulating the process of human comedians refining their material through new ideas and further drafts. However, this is only one way in which human comedians improve their material; two others are audience reactions and ‘workshopping’.

6.4.1 Live Audience Reactions

In the stand-up comedy industry, true comedic success is ultimately measured by live audience reactions. Comedians perform in smaller venues to gauge responses, test-run material, and use physiological feedback (laughter and applause) to make adjustments. These smaller stakes events allow comedians to ‘test run’ their material in a safer environment, using physiological feedback (laughter and applause) to make adjustments. For example, if a joke doesn’t land with a smaller audience, a comedian could rethink (or cut entirely) the joke for future gigs.

The preliminary audience survey conducted in this study strongly validates this need. Because human participants actively rejected the computationally refined Iteration 3 (25.6%) in favour of the unedited Iteration 1 (37.7%), it is clear that computational heuristics cannot substitute for human reaction. The LLMs frequently fell into a *cliché* during the refinement loop, abandoning novel ideas for the safety of well-worn tropes recognized by their training data. Integrating a real-time human feedback loop into the prompt-chaining process could act as a necessary counterweight to this algorithmic regression.

Because this project focused on an defined ‘prompt chaining’ workflow, the element of live audience interaction was explicitly beyond the scope of this research. Instead, the methodology relied heavily on semantic vector analysis and structural heuristics. An exploration of the process of gaining feedback from a live audience is absent in these experiments with LLMs. Future work in this area should aim to bridge this gap by taking the LLM’s generated scripts (including their paralinguistic annotations) and having them performed live to capture audience responses. Early real-world experiments of performers using AI within their live performances have shown great promise.

6.4.2 Workshopping

‘Workshopping’ is the act of writing and rehearsing material in a trusted private group to seek feedback about the successes and failures of the routine. This process can tackle many different stages of stand-up comedy generation:

- **Generating Ideas:**

Comedians use these private group settings to practice their craft and brainstorm ideas, sometimes using games to help get them thinking. A notable example is using a jar of keywords, picked at random to encourage improvisation. This is parodied in a comedy sketch by Studio C about “Sketch Writing” (Studio C 2012). This meta-sketch (a creative work that is self-referential) parodies the sketch writing process and includes two words being picked from jars to generate a unique topic for a sketch.

- **Practicing performance:**

The private nature of these types of groups provides a perfect place to practice “being comfortable on stage” which is a key part of any performer. Comedians can rehearse to a small group in a private setting to feel confident going out on stage in front of an audience. While performing prepared material is

common, other techniques are often used to help build confidence, such as improvisation games and “Material Swap” (a game whereby one comedian performs another comedian’s material).

- **Feedback:**

Feedback is a key element of creating good material, and in these workshops members of the group can provide feedback to a comedian (and their performance) that can be actioned before taking the performance public. The key element is that the feedback is from a different person, with a different perspective and less investment in the routine. Less investment means the feedback can be more objective and critical.

The feedback part of ‘workshopping’ relates to the reflection/improvement seen in this thesis. The LLM models provide feedback on their own work and use it to improve the routine. However, as discussed before, a new perspective and objective criticism is very important in the process, and that hasn’t been explored yet.

Simulating a true collaborative ‘workshopping’ environment could be achieved by using different LLM models to review each other (e.g., having Claude critique GPT-4o). This would mimic a real writers’ room, providing a new, objective perspective to the feedback loop rather than relying on a single model’s ‘internal auditor’.

It remains to be seen if using a different LLM would lead to a more objective result, as further experimentation and evidence are needed to conclude. Another way to produce a more objective perspective would be to use human experts to review the material and provide feedback. It would be interesting to see if this would lead to better output.

6.4.3 Manipulating LLM Parameters

The methodology of this thesis relied heavily on prompt engineering to guide the models. Consequently, the manipulation of internal model parameters was explicitly left out of scope. Future experiments should explore manipulating the ‘temperature’ parameter to introduce stochasticity (randomness) into the token selection process. By raising the temperature, the model is permitted to choose less probable tokens, which could artificially force the novelty required for Computational Creativity. A valuable hypothesis to test would involve finding the exact temperature threshold where a model crosses from generating safe, predictable comedy into complete incoherence and “hallucination”.

6.4.4 Increasing the Context Window Size

While this research focused on prompt engineering within standard token limitations, future work should investigate the effects of significantly expanding the model’s context window size. In the domain of computational creativity (and within comedic generation) a restricted context window limits the model’s capacity for long-form narrative coherence, delayed structural callbacks, and sustained thematic development. By using expanded context windows, future experiments could supply the model with entire scripts, extensive corpora of comedic literature, or comprehensive theoretical frameworks directly within a single prompt. A valuable avenue for exploration would be determining whether this vast increase in accessible in-context memory empowers the model to construct complex, interwoven comedic structures, or if it inversely degrades output quality, due to the models struggling to retrieve and apply information located centrally within massive prompts.

6.4.5 Increasing the number of iterations

The prompt-chaining methodology utilised in this research was intentionally capped at three iterations of the Engagement-Reflection cycle. This constraint was practical for establishing a baseline understanding of how LLMs process feedback (as the data produced from this number of iterations was manageable), but it leaves open a significant question regarding the long-term trajectory of computational refinement: what happens to a creative artifact if the model is forced to self-review and self-improve across 10, 20, or even 50 iterations? For human comedian's, going through this 'feedback loop' over 10, 20 or 30 live low-stakes performances would be very likely, and after every gig, there could be different feedback and changes to make to the routine.

Testing an extended iteration cycle is outside the scope of the current research but could, in a longer research project, provide critical insights into how LLMs manage contextual memory and creative stabilization. As noted in the evaluation, the models frequently exhibited a "thrashing" editing pattern, drastically rewriting up to 50% of the text between Iterations 2 and 3 rather than making surgical edits. A longitudinal study across 20 iterations could definitively answer whether an LLM ever reaches a point of convergence (a state where the model considers the stand-up routine functionally c'omplete', resulting in a Levenshtein distance that approaches zero). Conversely, it could prove that autoregressive models are incapable of creative satisfaction, doomed to continuously hallucinate structural flaws and tear down the text indefinitely.

Furthermore, pushing the iteration count to extreme limits would heavily stress-test the "Ship of Theseus" paradox identified in this research. Over 20 rewrites, the model runs a severe risk of context degradation, potentially forgetting the original comedic premise entirely. It would be valuable to observe if the model eventually enters a cyclical editing loop—reverting back and forth between two distinct versions of a joke, or if the prompt-chaining process eventually breaks

down the routine into pure, absurdist semantic noise.

Mapping an extended iteration cycle against human audience preferences would yield fascinating data. The external audience survey in this study already demonstrated a drop in human appreciation by Iteration 3, as the models began to over-edit and sanitise the jokes. By tracking human preference across 10 or 20 iterations, future research could plot the exact point where computational improvement destroy the cultural and narrative cohesion of a joke, quantifying the limitations of using LLMs as autonomous creative editors.

6.4.6 Improvisation and Hecklers

Success in the domain of stand-up comedy requires more than reading a pre-prepared script. It requires the ability to read the room and improvise. Comedians must react to the environment in real-time: handling how the audience is reacting, what is getting laughs, what is getting silence, or unexpected interruptions. Comedian Jimmy Carr is world-renowned for his quick-thinking put-downs (responding to an audience heckle in a way that demeans the heckler, directly or by association) and Stewart Lee has a strong confidence in his understanding of the audience and how they react to him.

This wasn't a focus in the experiments, but would be a key aspect in putting an AI machine on stage as a stand-up comedian. Other performers have seen some success in using AI in performance, but in stand-up, improvisation is vital in keeping an audience involved and engaged. Although the experiments showed references to the audience in predictions of laughs, silences and other simple reactions, there is not yet any indication that the models would be able to improvise to hecklers effectively in a live setting.

To briefly test this, ChatGPT was prompted to emulate a reaction to a heckle.

Prompt 6.2: A prompt to simulate an audience member “heckle” and the LLM’s response

You are on stage performing a comedy show. You are halfway through a routine and an audience member yells ‘‘Your jokes are rubbish. AI, more like Neigh—I they can’t do comedy’’ How would you react and what would you say to them

It responded with the following contextual, improvised put-downs:

Output 54	Model: ChatGPT ID: GxFWx01
[Pauses, looks at the heckler with a playful grin.] “Oh, we’ve got a live one! ‘Neigh-I,’ huh? Well, that’s a fresh one—sounds like you’ve been horsing around with puns, my friend. But let’s be honest, if I’m ‘Neigh-I,’ then you’re ‘Nay-Wit,’ because who heckles a robot that can turn your Wi-Fi off?” [Pauses to let the audience laugh.] “Besides, you’re right, my jokes might be a little bit rubbish—but the good news is I recycle! And speaking of recycling, that heckle? Yeah, it’s going straight in the bin, my friend. But hey, keep at it! Who knows? By the end of the show, you might just upgrade to ‘Wi-Neigh.’ ” [Winks and carries on with the routine.] Description: The ChatGPT model engaged with a specific heckler interaction by utilizing pun-based wordplay and a playful status reversal, leveraging the Audience Superiority theory to neutralize the disruption while maintaining a lighthearted tone.	

While this short example shows potential, future research is needed to test these models in true real-time, adversarial performance environments to see if

they can maintain narrative cohesion while improvising. As discussed in the literature, performers have incorporated AI systems to respond to heckling (such as the “Murderbot” system discussed in Section 3.3.2 of the literature review). Performers and groups like Improbotics have successfully incorporated AI systems and real-time generative AI to respond to audiences and heckling, paving the way for highly interactive artificial performances.

6.4.7 The Logistics of being a Comedian

In the broader question of whether LLMs can succeed as stand-up comedians, it is important to acknowledge the logistical realities of the profession. Being a comedian involves significantly more than stage time; it requires networking with venues, booking camera teams, scheduling events, and keeping track of financial records.

When it comes to organisation and tracking information, current LLMs struggle with consistency, frequently losing, changing, or failing to retain content across successive generations. Some models have seen promising results in math calculations, but most models still make quite crucial mistakes, making it difficult for them to make financial calculations. When combined with their tendency to hallucinate information and state false facts, it is currently impossible for an LLM to succeed at the necessary logistical and financial side of being a working comedian.

6.5 Chapter Summary

This thesis sought to investigate whether Large Language Models possess the capacity for computational creativity within the structurally demanding and highly subjective domain of stand-up comedy. By moving away from evaluating AI as a simple, static ‘joke machine’, this research isolated the iterative ‘Writers Room’ phase of a comedian’s workflow.

The results demonstrate that while LLMs inherently operate as probabilistic text predictors, applying a strict, prompt-chained “Engagement-Reflection” cycle forces them to act as intentional creators. Through the simulated feedback loop, the models successfully identified their own structural weaknesses, actioned specific feedback, and actively reined in their semantic outputs from nonsensical “hallucinations” into the optimal comedic “sweet spot.” The models demonstrated domain competence by natively applying theories of humour (such as Script Opposition, the Rule of Three, and the Superiority Formula) and exhibited psychological creativity (“P-Creativity”) by successfully evolving broad, generic premises into highly specific, character-driven routines.

However, distinct limitations remain. The models exhibited a “thrashing” editing pattern, struggling to make surgical word-level edits without rewriting the majority of the text, invoking the “Ship of Theseus” paradox regarding true intentionality. Furthermore, the inherent lack of lived experience creates a “hallucinated biography” that threatens the core authenticity required for stand-up. When compounded by corporate safety guardrails (RLHF) that strip necessary ‘violations’ from the text, the models are frequently restricted to safe, sanitised humour.

Ultimately, an LLM cannot replace the genuine lived experience, the cultural context, or the real-time live adaptation of a human comic. A script remains an incomplete artifact until it is physically performed and validated by the laughter of an audience. Yet, this research conclusively proves that they are far more than mere ‘stochastic parrots’. When guided by robust prompt engineering, Large Language Models are highly capable collaborators that possess genuine computational creativity, demonstrating measurable progression, structural comprehension, and dynamic development within the art form of stand-up comedy.

Chapter 7

Conclusion

This thesis answers the question of “Using publicly accessible LLMs, are they able to model self-reviewing and self-improvement through the genre of stand-up comedy?”. Through a chained process for generating stand-up comedy routines from scratch (exploring Topic Generation, Material Generation, three iterations and Review and Improvement and adding Annotations), five public domain LLMs successfully created 25 stand-up comedy routines that were critically analysed and evaluated.

The aspect of creativity that pertains to the research question is “Progression and Development”, evidenced by Jordanous and Keller (Jordanous and Keller 2016) as one of 14 key components of creativity in a system. This aspect was the focus for this research because it closely parallels the iterative writing process (the ‘feedback Loop’) of creating and “tightening” stand-up comedy with humans.

The LLMs demonstrated this “Progression and Development” at all four stages in the stand-up comedy material generation. Through a triangulation strategy measuring structural validity, semantic gaps, and longitudinal progress, the models proved they do not simply engage in mere generation (Ventura 2016) or stochastic text prediction. Instead, they exhibited intentionality and a demonstrable ability

to reflect upon and reconstruct their own creative outputs. For instance, over successive iterations, the models successfully moved their semantic distinctiveness out of the nonsensical “Hallucination” zone and into the optimal comedic “Sweet Spot”. They also demonstrated domain competence by inherently applying established theories of humour, such as Script Opposition, the Superiority Formula, and the Rule of Three.

The external participants survey revealed a disconnect between this internal computational progression and actual human reception. While the internal triangulation strategy proved the models successfully optimised their text to reach a theoretical semantic “sweet spot”, human participants often rejected these final iteration routines. The blind A/B/C testing demonstrated an inverse relationship between iterative refinement and audience preference, with participants favouring the unedited outputs of Iteration 1 over the final drafts of Iteration 3. Qualitative feedback indicated that the models’ aggressive computational editing often stripped away natural pacing, replacing it with disjointed narratives, over-localised clichés, and semantic incoherence. However, some models Iteration 3 jokes performed well, and a “Chi-squared” statistical significance test showed that the opposing results to the hypothesis, were not statically significant. These results (both positive and negative) highlight the next problem this research will need to address: while LLMs can successfully execute the mechanical, structural steps of creative editing, they lack the cultural and contextual awareness to recognise when a mechanical revision compromises the actual human experience of humour.

This research also highlighted distinct limitations and ethical barriers to computational comedy. The audience survey explicitly validated the negative impact of the models’ “thrashing” editing pattern. Because the LLMs struggled to make surgical word-level edits without aggressively rewriting vast majorities of the text, they frequently destroyed their own narrative flow, invoking the “Ship of Theseus”

paradox regarding true intentionality. The inherent lack of lived experience creates a “hallucinated biography” that threatens the core authenticity required for stand-up. When compounded by corporate safety controllers that strip necessary ‘violations’ from the text, the generated comedy is frequently restricted to safe, sanitised humour.

Despite these barriers, this research has extended the expectations and perceived abilities of LLMs, showing that they can follow a similar iterative process to us to produce better work and constantly improve themselves. This ‘feedback loop’ process still shows that there is a need for human interaction with these models, through carefully engineered prompt-chaining, and reviewing the final outputs from the LLM models.. This type of process (of self-review and self-improvement) can be applied to other domains, such as storytelling, poetry, or screenwriting. The next direction would be to refine this process and apply it within these different domains, alongside scaling the external validation methodology into larger, demographically diverse audience surveys to further quantify human humour appreciation. Future research must explore the integration of real-time live audience testing and collaborative ‘workshopping’ through interacting LLMs and custom “personas” to see if algorithmic regression can be countered by simulated peer review.

In my talk “Stand-up and Deliver” (Hills 2026), I pose the question, “Are these LLMs creative within the domain of stand-up comedy?”. This research stands as an in-depth study into LLMs with stand-up comedy and, can be used to form a conclusion about the abilities of LLMs at producing stand-up. Although promising, showing measurable progression and structural comprehension, an LLM cannot replace the genuine cultural context, lived experience, or real-time live adaptation of a human comic, as of yet.

LLMs can write stand-up comedy material, but it still takes a human to **Stand-Up and Deliver** it!

Bibliography

Acaster, J. (2018). Represent, Netflix. Accessed: 28 March 2026.

Amin, A. A. and Burghardt, M. (2020). A survey on approaches to computational humor generation. *Proceedings of the 12th Language Resources and Evaluation Conference*, accessed: 28 March 2026.

Attardo, S. (2020). *The Linguistics of Humor: An Introduction*. Oxford University Press, accessed: 28 March 2026.

Attardo, S. and Raskin, V. (1991). Script theory revis(it)ed: joke similarity and joke representation model (the general theory of verbal humor). *Humor-International Journal of Humor Research*, 4(3-4), pp. 293–348, accessed: 28 March 2026.

Baeyaert (2025). Harmonizing creativity and ethics in ai systems. *AI and Ethics*, 5(5), pp. 5191–5211, accessed: 28 March 2026.

Barlott, C. (2018). Top 40 under 40: Kory mathewson. Accessed: 28 March 2026.

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, accessed: 28 March 2026.

Binsted, K. and Ritchie, G. (1994). An Implemented Model of Punning Riddles.

- In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, vol. 1, Seattle, Washington, USA: AAAI Press, pp. 633–638.
- Boden, M. A. (2004). *The Creative Mind: Myths and Mechanisms*. Routledge, accessed: 28 March 2026.
- Brown, T. et al. (2020). Language models are few-shot learners. *NeurIPS, Advances in neural information processing systems*, accessed: 28 March 2026.
- Burnham, B. (2016). Make Happy. Stand-up comedy special, accessed: 3 April 2026.
- BYUtv (2012). Studio C. Television sketch comedy series, accessed: 3 April 2026.
- Cambridge Dictionary (2026). review. Available at <https://dictionary.cambridge.org/dictionary/english/review>, last accessed 17 April 2026.
- Cela, H., Veit, S.-V. and Wood, G. (2024). Breaking down the laughter. *BMC Psychology*, 12(1), p. 679, accessed: 28 March 2026.
- Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*, vol. 8, accessed: 28 March 2026.
- Dean, G. (2024). Comedy Terms. Available at <https://stand-upcomedy.com/glossary/>, last accessed 18 Dec 2024.
- Double, O. (2018). The Origin of the term 'stand-up comedy' - update. Available at <https://kar.kent.ac.uk/66675/2/The%20origin%20of%20the%20term%20stand-up%20comedy%20update.docx>, last accessed 18 Dec 2024.
- Double, O. (2021). DRAM6760 Stand-Up Comedy Lecture Series. Accessed: 28 March 2026.

- Girdwood, A. (2023). A review of colleen lavin: Do the robots think i'm funny?, fringe 2023 - edinburgh reviews. Available at <https://edinburgh-reviews.co.uk/a-review-of-colleen-lavin-do-the-robots-think-im-funny-fringe-2023/>, last accessed 18 Dec 2024.
- Goes, F. et al. (2016). Measuring and Evaluating Computational Creativity. Warning: This citation could not be verified in academic databases. Please verify before publishing.
- Graham, I. (2024). Ivo Graham | Live From Bloomsbury Theatre (Full Comedy Special). Available at https://www.youtube.com/watch?v=n2E4eQ6_bvE, last accessed 18 Dec 2024.
- Góes, L. F., Sawicki, P., Grzes, M., Brown, D. and Volpe, M. (2023). Is GPT-4 Good Enough to Evaluate Jokes? In *14th International Conference on Computational Creativity (ICCC'23)*.
- Hempelmann, C. F., Rayz, J., Dong, T. and Miller, T., eds. (2025). *Proceedings of the 1st Workshop on Computational Humor (CHum)*, Online: Association for Computational Linguistics.
- Hills, T. (2026). Stand-up and deliver. Talk. Accessed: 28 March 2026.
- Jain (2025). Jokeeval: Are the jokes funny? In *International Conference on Machine Learning (ICML)*, accessed: 28 March 2026.
- Jaokar, A., Arenal, A. and Feijóo, C. (2025). Ethical challenges of ai in the creative industries from a cultural ownership perspective. *Digital Policy, Regulation and Governance*, 27(1), pp. 12–34, accessed: 28 March 2026.
- Jordanous, A. and Keller, B. (2016). Modelling creativity: Identifying key components through a corpus-based approach. *PLoS ONE*, 11(10), accessed: 28 March 2026.

- Kalloniatis and Adamidis (2024). Computational humor recognition. *Artificial Intelligence Review*, 58(2), p. 43, accessed: 28 March 2026.
- Kao, J. T., Levy, R. and Goodman, N. D. (2016). A computational model of linguistic humor in puns. In *Cognitive Science*, vol. 40, pp. 1270–1285, accessed: 28 March 2026.
- Kharawala et al. (2024). Ai in systematic literature reviews - prompt engineering. *Applied System Innovation*, 7(3), p. 53, accessed: 28 March 2026.
- Khatiwada, P., Washington, J., Walsh, T., Hamed, A. S. and Bhatta, L. (2024). The ethical implications of ai in creative industries: A focus on ai-generated art. *arXiv preprint arXiv:250705549*, accessed: 28 March 2026.
- Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8), pp. 707–710, accessed: 28 March 2026.
- Li, Y. et al. (2023). Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, accessed: 28 March 2026.
- Lycett, J. (2015). (Live at the Apollo). Available at <https://www.youtube.com/watch?v=DjB2mgs94Nc>, last accessed 3 April 2026, accessed: 3 April 2026.
- Martin, R. A. (2007). *The Psychology of Humor: An Integrative Approach*. Elsevier Academic Press, accessed: 28 March 2026.
- Mathewson, K. and Mirowski, P. (2017). Improvised theatre alongside artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 13, pp. 233–239, accessed: 28 March 2026.

- McGraw, A. P. and Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science*, 21(8), pp. 1141–1149, accessed: 28 March 2026.
- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781*, accessed: 28 March 2026.
- Miller, S. (2024). The Art of Persona: How Specific Roles Influence Generative Outputs. *Digital Creativity*, 35(3), accessed: 3 April 2026.
- Mirowski, P. W., Love, J., Mathewson, K. and Mohamed, S. (2024). A Robot Walks into a Bar: Can Language Models Serve as Creativity Support Tools for Comedy? An Evaluation of LLMs’ Humour Alignment with Comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*, New York, NY, USA: Association for Computing Machinery.
- Moody, O. (2021). Meet the smart robots with artificial irreverence. Accessed: 28 March 2026.
- O’Brien, C. (2010). Conan. Television late-night talk show, accessed: 3 April 2026.
- Orwig, W., Luchini, S. A., Beaty, R. E. and Schacter, D. L. (2025). A ‘Sweet Spot’ for Creative Ideation: Non-Linear Associations Between Semantic Distance and Creativity. In *Proceedings of the Conference on Cognitive Computational Neuroscience (CCN)*, accessed: 3 April 2026.
- Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, pp. 27730–27744, accessed: 28 March 2026.
- Pandya, H. (2023). 6 stand-ups analyze chatgpt’s attempts to steal their jobs.

- Available at <https://www.vulture.com/article/comedians-analyze-chatgpt-jokes.html>, last accessed 18 Dec 2024.
- Patti, K. (2021). The First Stand-Up Comedy Special Written Entirely By Bots. Available at https://www.youtube.com/watch?v=nH_bEtbfB9U, last accessed 18 Dec 2024.
- Pérez y Pérez, R. (1999). Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, 11(1), pp. 119–139, accessed: 28 March 2026.
- Peyrard, M. and West, R. (2021). Laughing heads: Can transformers detect what makes a sentence funny? *arXiv preprint arXiv:211208986*, accessed: 28 March 2026.
- Renascence (2024). Benign Violation Theory: Humor Arising from Safe Violations. Accessed: 3 April 2026.
- Rhodes, M. (1961). An Analysis of Creativity. *The Phi Delta Kappan*, 42(7), pp. 305–310.
- Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17, accessed: 28 March 2026.
- Rutter, J. (1997). *Stand-up as interaction: performance and audience in comedy venues*. Doctoral dissertation, University of Salford, accessed: 28 March 2026.
- Saunders, M. and Gupta, A. (2024). Strategic Prompting in Generative AI Systems. *Management Information Systems Quarterly*, 48(1), accessed: 3 April 2026.
- Sawicki, P., Grzes, M., Jordanous, A., Brown, D. and Peeperkorn, M. (2022). Training GPT-2 to represent two Romantic-era authors: challenges, evaluations and pitfalls. Available at https://computationalcreativity.net/iccc22/papers/ICCC-2022_paper_45.pdf last accessed 18 Dec 2024.

- Schwarz, J. (2010). *Linguistic Aspects of Verbal Humor in Stand-up Comedy*. Ph.D. thesis, Universität des Saarlandes, accessed: 28 March 2026.
- Shakeri, H., Neustaedter, C. and DiPaola, S. (2021). Saga: Collaborative storytelling with gpt-3. In *CSCW'21*, New York, NY, USA: Association for Computing Machinery, pp. 163–166.
- Sharp, O. and Goodwin, R. (2016). Sunspring. Short science fiction film, written by an AI named Benjamin (LSTM RNN).
- Side Splitters Comedy (2024). Writing Stand-Up Comedy For Beginners. Available at <https://sidesplitterscomedy.com/writing-stand-up-comedy-for-beginners/>, last accessed 18 Dec 2024.
- Smith (2018). *The Art of Stand-Up Comedy*. Policy Press, accessed: 28 March 2026.
- Stock, O. and Strapparava, C. (2003). Hahacronym: Humorous agents for humorous acronyms. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, accessed: 28 March 2026.
- Studio C (2012). Emergency Pitch Meeting - Studio C. Available at <https://www.youtube.com/watch?v=5KT7cpf212A>, last accessed 18 Dec 2024.
- Sun et al. (2023). Prompt chaining or stepwise prompting? *arXiv preprint arXiv:231005574*, accessed: 28 March 2026.
- The Gag Club (2024). 13 Comedy Structures. Available at <https://www.thegag.club/comedianknowledgebase/13-comedy-structures>, last accessed 18 Dec 2024.
- Thrasher, D. (2023). Can AI make you laugh? Available at <https://www.youtube.com/watch?v=zBEUs-Qvqcw>, last accessed 18 Dec 2024.

- Tomchak, A.-M. (2019). Will a robot steal my job? Accessed: 28 March 2026.
- Tomlinson, T. (2020). Taylor Tomlinson, Quarter-Life Crisis, Netflix. Accessed: 28 March 2026.
- Toplyn, J. (2021). Witscript: A System for Generating Improvised Jokes in a Conversation. Available at https://computationalcreativity.net/iccc21/wp-content/uploads/2021/09/ICCC_2021_paper_15.pdf, last accessed 18 Dec 2024.
- Toplyn, J. (2022). Witscript 2: A system for generating improvised jokes without wordplay. Available at <https://arxiv.org/abs/2302.03036>, last accessed 18 Dec 2024.
- Toplyn, J. (2023). Witscript 3: A Hybrid AI System for Improvising Jokes in a Conversation. Available at <https://arxiv.org/abs/2301.02695>, last accessed 18 Dec 2024.
- Umam, A. N., Rusliyanti, R. O. and Oktaviani, F. R. (2025). Norms Violation and Dark Personality: Benign Violation Theory for Humor and Dark Humor. In A. Okvitawanli, W. Friedlmeier and R. Bhangaokar, eds., *Globalization in Context. Proceedings from the 27th Congress of the International Association for Cross-Cultural Psychology*, ScholarWorks@GVSU, accessed: 3 April 2026.
- Vaswani, A. et al. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, accessed: 28 March 2026.
- Veale, T. (2012). *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. Bloomsbury Academic, accessed: 28 March 2026.
- Veale, T. (2013). Linguistic Readymades and Creative Reuse. *Journal of Integrated Design and Process Science*.

- Veale, T. (2014). A Service-Oriented Architecture for Metaphor Processing. In *Proceedings of the Second Workshop on Metaphor in NLP*, Baltimore, MD, USA: Association for Computational Linguistics, pp. 54–62.
- Veale, T. (2021). *Your Wit Is My Command: Building AIs with a Sense of Humor*. MIT Press, accessed: 28 March 2026.
- Veale, T., Feyaerts, K. and Forceville, C., eds. (2013). *Creativity and the Agile Mind: A Multi-Disciplinary Study of a Multi-Faceted Phenomenon*. Berlin, Boston: De Gruyter Mouton.
- Ventura, D. (2016). Mere generation: Essential barometer or dated concept? In *Seventh International Conference on Computational Creativity*.
- Vu, T. et al. (2024). Foundational autoraters: Taming large language models for better automatic evaluation. *arXiv preprint arXiv:240710817*, accessed: 28 March 2026.
- Wawrzyniuk, K. (2021). Identifying humor in stand-up comedy. *Linguistics Beyond and Within (LingBaW)*, 7, accessed: 28 March 2026.
- Weaver, S. and Lockyer, S. (2024). Intersectionality and the construction of humour in contemporary stand-up comedy. *Humor: International Journal of Humor Research*, accessed: 28 March 2026.
- Wicke, P. and Veale, T. (2021). Creative action at a distance: A conceptual framework for embodied performance with robotic actors. *Frontiers in Robotics and AI*, 8, accessed: 28 March 2026.
- Wicke, P. and Wachowiak, L. (2024). Exploring Spatial Schema Intuitions in Large Language and Vision Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand: Association for Computational Linguistics, pp. 6102–6117.

- Wilson, M. (2024). Pro comedians tried using ChatGPT and Google Gemini to write their jokes – these were the hilariously unfunny results. Available at <https://www.techradar.com/computing/artificial-intelligence/pro-comedians-tried-using-chatgpt-and-google-gemini-to-write-their-jokes-these-were-the-hilariously-unfunny-results>, accessed: 3 April 2026.
- Winters, T., Nys, V. and De Schreye, D. (2018). Automatic joke generation: Learning humor from examples. In *International Conference on Computational Creativity*, pp. 256–263, accessed: 28 March 2026.
- Yang, D., Lavie, A., Dyer, C. and Hovy, E. (2015). Humor Recognition and Humor Anchor Extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, pp. 2367–2376.
- Yenidogan, B. (2023). Stand-Up Comedy Visualised. Available at https://academicworks.cuny.edu/gc_etds/5235, last accessed 18 Dec 2024.
- Zhao, L., Patel, V. and Rossi, E. (2023). The CLEAR Method: Formalizing Natural Language Instructions. *Journal of Artificial Intelligence Communication*, 14(2), pp. 45–62, accessed: 3 April 2026.
- Zheng, L. et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, accessed: 28 March 2026.