Interactive Storytelling with ChatGPT: An Explorative study

University of Turku
Department of Computing
Master of Science Thesis
Interaction Design
December 2025
Aleksi Torri

UNIVERSITY OF TURKU
Department of Computing

Aleksi Torri: Interactive Storytelling with ChatGPT: An Explorative study

Master of Science Thesis, 48 p., 2 app. p.
Interaction Design
December 2025

Since the release of general-purpose generative AI tools such as OpenAI's ChatGPT and Google's Gemini, there has been a large, sometimes exaggerated, amount of publicity on the capabilities of these contemporary language models. Through the lens of interactive storytelling, these capabilities narrow down to a model's ability to generate narrative content, and its ability to maintain narrative cohesion.
The focus of this thesis is on ChatGPT 4o Mini's storytelling abilities using a framework we designed, called Dungeon Master. To test this AI model's abilities within this framework, we used a pre-written story outline as the base narrative, and created three prompt models with different levels of allowance for the generative model to modify it: Lenient, Medium, and Strict. We conducted ten tests per prompt model and analysed them based on the number of divergences that occurred in each instance of a story.
Our results indicate that this model of ChatGPT manages to generate a coherent and narratively consistent storyline for the most part with some minor caveats. The three prompt models we designed and tested had a minimal impact on the way the narratives evolved.

Keywords: ChatGPT, Interactive Storytelling, Interactive narrative, prompt engineering

Contents

List of Figures

1 Introduction

Contemporary role-playing games, or RPGs for short, give players an ever-increasing level of control over the game world. In games like *Divinity II: Original Sin*, *Star Wars: The Old Republic*, and *Baldur's Gate 3*, the game world is already highly engaging, with hundreds of interactable elements per level. Despite this high level of interactability when it comes to the game world and the physics engine, an issue that seems to frustrate some players constantly has been the illusion, or lack thereof, of choice in the storyline. This issue arises from multiple sources, including the narrative paradox and the cost required to create these complex and winding storylines.

This thesis is an exploratory study into the use of generative artificial intelligence (AI) models in interactive storytelling, it focuses on the possibility of managing a dramatic story arc using a drama manager. Although it does not explicitly focus on RPGs, it focuses on the general use of generative AI within computer role-playing systems, using the creation of narrative content for RPGs as a guiding principle for the research direction. The goal of this thesis is to explore what can be achieved with minimal resources and how effectively general-purpose generative AI systems can be used to create interactive narratives while acknowledging that a purpose-built system would likely be a more effective option.

We will be using ChatGPT, a generative AI system developed by OpenAI, to create interactive narratives and explore how well it can be used in this context.

Specifically, we will be using the 4oMini model (released July 18, 2024), which was the version available to free users of ChatGPT at the time of running the tests. We will explore various testing frameworks (hereafter used interchangeably with approaches) that may be used with ChatGPT in interactive narrative creation, then focus on one of these frameworks in more detail. This approach, which we have dubbed the Dungeon Master, will be tested with three different model variants, each providing a distinct level of control over the story to the generative AI: the Lenient, Medium, and Strict models.

We predict that the Medium model will give the best results, as it gives ChatGPT enough freedom to create a compelling story while still keeping the story on track. We also propose that the Lenient model will be the most chaotic, as it gives the AI too much freedom to create a story that is not coherent and does not follow any dramatic structure. The Strict model, on the other hand, will follow the outline of the story too closely, leading to a loss of replayability and a lack of immersion in the story.

In this thesis, we also explore how well ChatGPT will adhere to genre restrictions, in our case, a historical setting. We will be tracking any ahistorical or fantastical elements that ChatGPT or the player introduces into the story. We made this choice because it allows us to verify the accuracy of historical facts easily. Here, we predict that all three models will perform similarly, so in other words, none of them will introduce any such elements into the story, as long as the player does not explicitly ask for them.

At the outset of this thesis, we defined AI hallucinations as nonsensical responses, those that do not align with the internal narrative logic or sentences that are simply gibberish, rather than focusing on the misrepresentation or false information which, according to our research, tends to be the more commonly accepted definition within the field of computer science. In addition, some research highlights that that the

use of the term “hallucination” can have unintended, harmful, associations towards mental health problems [1]. Since interactive storytelling and narratology is not about factual correctness but rather internal logic, and to avoid stigmatizing mental health issues, we needed new terms that better capture these ideas. This is how we ended up grouping genre adherence with concepts we dubbed as “event sequential coherence” and “semantic coherence” into “narrative coherence”.

This thesis is divided into six chapters. The second chapter, which follows this one, introduces you to the important terminology used within this work and explains the basics of the technology used in later chapters. The third chapter explores the ethics of using artificial intelligence, and specifically generative language models, in the creation of narratives. In the fourth chapter you will find both the background research that this work is based on, and the conceptual frameworks that we developed during this thesis. This chapter also contains an explanation of the storyline utilized in the following study. The fifth chapter explains the methods we used to carry out the tests and lays out the observations made during the preliminary testing phase. In the sixth chapter you will find our analysis of the exploratory study that was done in conjunction with this thesis work. The final chapter is dedicated to a more freeform conclusion and discussion on the results and how this research could be continued in the future.

2 Terminology

In this chapter, we will explore the relevant terminology to this work. Clarifying these concepts not only improves the readability of this thesis, but also provides a common context for these terms to readers. While some concepts, like the narrative paradox, are already well established in the field of interactive storytelling, others are more specific to this work, like the distinction between the three storytelling agents. Furthermore we will delve into external tools that are pertinent in the context of this work, like ChatGPT.

By examining the definitions in interactive storytelling, the roles of the author and the storyteller, and the distinctions between types of storytellers, we aim to create a solid foundation to the work that follows. These terms are essential, as they influence the dynamics of audience influence and narrative development shaping the way we understand the emergence of interactive narratives and thus, how we aim to improve these dynamics through this work.

2.1 Interactive storytelling

While a clear line cannot be drawn as to what is included within interactive storytelling in general terms, as many would argue that, for example, some forms of advertisements are considered this type of narratology, in this thesis, we will use a similar definition as Bostan and Marsh [2]. Thus, for the purposes of this study, interactive storytelling is a sub-genre of storytelling where the audience of the story

is able to influence the course of the story in some way. These types of narratives range from simple text adventure games and role-playing games all the way to interactive movies and theatre. In this way, it is a broad term that encompasses various types of storytelling. It is both an intriguing and complex topic due to the human factor. No single story will be the same, since no audience is the same. With the advent of Large Language Models such as ChatGPT and Gemini, which are capable of generating an extensive amount of coherent natural language text, there are numerous avenues of research within this field.

Interactive narratology tells us that an interactive narrative develops from the interaction between three roles: the author, the drama manager, and the audience. In this, we divide the drama manager into three additional subcategories depending on how much control the author or the audience has over the story. We have the author-centric model, the player-centric model, and the hybrid model. In general, it is understood that the choices are between the author-centric and player-centric models, with the hybrid model being practically impossible to implement due to its nature as a compromise between the two extreme models [3]. In this thesis, we refine this classic model by distinguishing between the storyteller (the entity performing the narrative) and the drama manager, or the system-level controller. While the term drama manager often overlaps with our definition of a rule-based storyteller, maintaining this distinction allows us to analyse generative AI in terms of narrative performance rather than system logic alone. This new model is illustrated in Figure 2.1

2.2 Author

We will use the term author to refer to the person or piece of technology that creates the story. To date, this role has predominantly been fulfilled by humans, even in interactive digital narratives (IDN). However, with the emergence of AI, and es-

pecially LLMs that are able to generate large amounts of text data, we propose that it is advantageous to consider these models as potential authors. By broadening our understanding of authorship to include AI models, we can explore new avenues of narrative creation and examine the implications of AI-generated content on storytelling.

2.3 Storyteller

In this thesis, we will be using the term storyteller for the person or piece of technology that tells the story. This role can be fulfilled by a human, a dedicated software application, or a generative AI system. The storyteller is the entity that presents the author-written narrative to the audience in a captivating and engaging way, employing such techniques as pacing, tone and dramatic emphasis. While the author may fulfil the role of storyteller, this is not a requirement. This distinction is important in computer interactive narratives, since in this field, a dedicated piece of software (drama manager) often fulfils this role. We broaden this to include generative AI in its current form, enabling us to study its use in interactive narratology as a tool for making pseudo-creative decisions.

2.3.1 Rule-based storytellers

Rule-based storytellers are a type of storyteller that we define in order to differentiate between different types of storytellers. Rule-based storytellers would typically be a piece of software, but a human could also be a rule-based storyteller, though this would require not using any creativity when telling the story. A rule-based storyteller follows a set of rules, usually the literal outline of a story. It is unable to create new content and regurgitates the story as written by the author and thus, these storytellers are commonly found in mediums such as computer role-playing games

(CRPGs), where the narrative is predefined and it isn't performed by a human.

2.3.2 Creative storytellers

The creative storyteller is an entity that can use its creativity to create new content, deviate from the original story at any point, modify the story, and create new characters, locations, or events based on audience input, at will. We determine for the purposes of this thesis, that the only entities that can be categorized as creative storytellers, are humans, since no other entities are known to possess creativity, though there are claims that gen AI can be creative. For further discussion on these claims, refer to Chapter 3 on the ethics of using gen AI in storytelling.

2.3.3 Generative storytellers

Generative storytellers are a fairly new type of storyteller that, while lacking creativity, can mimic it in certain situations and with specific prompts. As the name suggests, these are strictly a category of storytellers based on the use of generative AIs. Generative storytellers are able to create new content on demand, but this new content is not new in the sense that it is novel, but rather that it is a recombination of an existing body of literature that the LLM has been trained on.

Since this thesis focuses on the storytelling and creative capabilities of ChatGPT, and since we have excluded gen AI's from the creative category due to a lack of proof in these models possessing creativity, we have created this category, as we assert that these tools are distinct enough from rule-based storytellers.

2.4 Player

In this thesis, we will be using the term *player* to identify the person who is experiencing the story. Although other terms, such as *reader* or *user*, could be used, we

contend that these alternatives do not encapsulate the nature of the interaction in a satisfactory way. The term *reader* does not capture the interactive element of the person's experience, while *user* is a term broadly used in the context of software for a person interacting with a program. We assert that it is important to distinguish between the broader concept of interaction within software, and the specific engagement characteristic found in interactive narratives. In addition, the act of interacting with such a narrative is more closely associated with the act of play and, thus, more closely linked with the concept of a player.

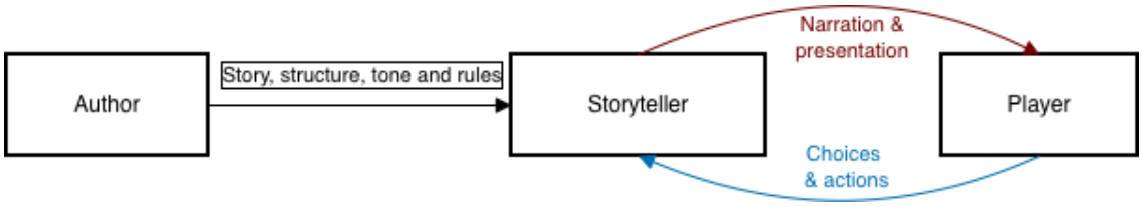

Figure 2.1: Relationship between the author, storyteller and player

2.5 Artificial intelligence

Artificial intelligence (AI) is a broad term used to denote the technologies that attempt to mimic human learning, reasoning or comprehension[4]. Most importantly for this thesis, these technologies include two frameworks of interest to us. The first approach is one based on relations between concepts defined by human-crafted rules, known as deterministic Natural Language Processing (NLP), or symbolic AI models. The second approach relies on predicting the next token or word in a sequence forming the basis of generative NLP models.

2.5.1 ChatGPT

ChatGPT is an AI framework developed by OpenAI[1]. It is a variant of the generative pre-trained transformer (GPT) models, which are NLP models trained on a large amount of text data. GPT models do not have any "intelligence" in a colloquial sense, as in, there is no clear evidence to believe that such a model understands the input it receives or the output it generates, rather, it predicts what the output would be based on the input [5]. In essence, such a language model is based upon a neural network that receives as input the previous messages, and it composes as output what is most likely to be the following words to the sequence inputted. While saying that it simply selects words one after another is slightly reductive, this explanation is not inaccurate in the most simplest of terms. Given a prompt, it determines the likelihood of a certain word following the previous ones, and then selects one word from a pool of most likely words. In this way, it can create whole sentences and paragraphs.

GPT models have a temperature parameter, which in human terms could be explained as it's "creativity" setting. According to the official OpenAI API documentation, this parameter is set between a range of 0–2 [6]. In this case, a low temperature, like 0.2, means that the model will act in a more deterministic manner, while a high temperature, like 1.7, means that the model will act with increased randomness [7]. While OpenAI has not published the default temperature parameter for the ChatGPT interface, it is generally assumed that it is set to a moderate value.

ChatGPT is a chatbot-style web interface (pictured in Figure 2.2), which works with a fine-tuned version of a GPT model. This model was provided additional data in the form of conversations, wherein the human AI trainers played both sides. Through this type of reinforcement learning (RL), OpenAI was able to create a version of their GPT models that mimicked the way a human would reply to another

[1]https://openai.com/about

human within a text chat. [8]

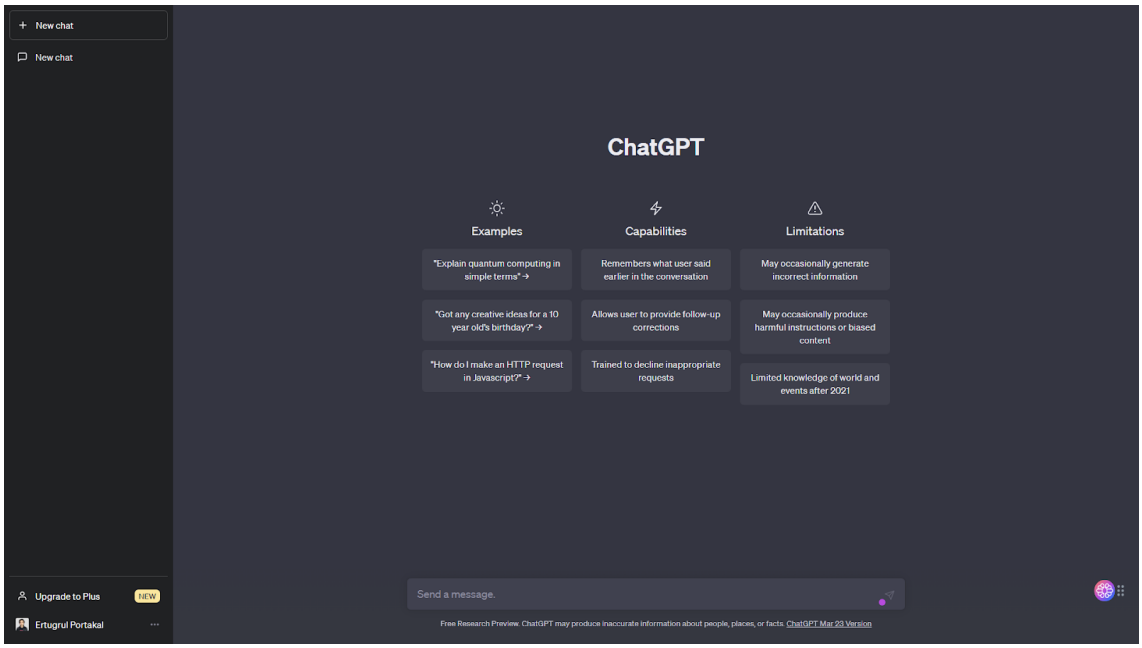

Figure 2.2: The ChatGPT interface

2.6 CLIPS

CLIPS, or C Language Integrated Production System, is a tool that has been developed by NASA for creating expert systems. An expert system uses a set of rules to make decisions. The current version of CLIPS is independently developed and maintained by Gary Riley [9]. In addition to expert systems, CLIPS can be used for a variety of tasks, such as text adventure games, which is why we are interested in it in this thesis. [10]

2.7 Drama manager

A drama manager is the story engine of most interactive storytelling architectures, it is the part of an interactive storytelling model that is responsible for managing

the story through so-called "story beats" [2].

In this thesis, we envisioned such a "drama manager" as a deterministic algorithm that would be a liaison between ChatGPT and the player. This algorithm would be responsible for guiding ChatGPT in the right direction, and in some of our approaches, would be responsible for evaluating the quality of the story. In the scope of this thesis, we will not develop any such algorithm, but we will simulate its functionality in the tests we conduct.

2.8 Narrative paradox

The narrative paradox is a term used to describe the tension between the need for a coherent story and the need for player agency [11], [12]. It is the inherent challenge when creating interactive digital narratives that emerges from the differing goals of the actors involved within the creation of these stories. The author wants to create a coherent narrative that captures the player's attention while the player wants a level of control and agency over the course of the narrative.

This problem has been studied for a long time, and there are many approaches, but they can generally be divided into three categories: The author-centric model, the player-centric model and the hybrid-model [3]. As generative models are introduced to the field of interactive storytelling, we are confronted with new challenges within the narrative paradox, as these models may erode the narrative coherence of a story.

3 Ethical considerations

In this chapter, we will discuss the ethics of using generative AI in authoring IDNs. The rise of generative AI tools, such as OpenAI's ChatGPT, has raised discussion on the ethics of using such tools in many fields, especially in artistic and creative fields. Since interactive storytelling is a form of creative writing, we believe that it is important to discuss the ethics of using these tools in our field. Most of this section is based on "Centering the Human: Digital Humanism and the Practice of Using Generative AI in the Authoring of Interactive Digital Narratives". [13]

3.1 The Digital Humanist approach

The Digital Humanist approach argues that the use of generative AI should be seen as a tool that can enhance the human creative process, rather than replace it. This perspective emphasizes the importance of human authorship and the valuation of human creativity in the face of technological advancements. It argues that the human author and the generative AI cannot share authorship, and that the human author is the only one who can be considered the author of the work.

3.2 The Post-humanist approach

Post-humanism argues that the human author and the generative AI should be seen as co-authors of the work. This perspective challenges the traditional notion of

authorship, which has historically been associated in a humanist framework wherein only a human may claim to be the author of a work.

3.3 The Transhumanist approach

The Transhumanist approach argues that the human author and the generative AI would, rather than being co-authors, be seen as one entity that is considered the author of the work. In other words the Gen AI and the human would combine into one to create the entity that is considered the author of the work.

3.4 Discussion

While we acknowledge the importance and value of human creativity, we also believe that it is important to consider the player experience and the co-authorship of certain elements of a story with the generative AI. For instance, in a dialogue, the player may want to ask a non-player character (NPC) a question that the author has not anticipated. In this case, the generative AI can be seen as a co-author, as it might be able to generate a response that is consistent with the story and the character's personality. This perspective allows for a more flexible and dynamic storytelling experience, where the player can interact with the story in a more meaningful way.

Considering these approaches, we believe that our approach is more aligned with the Post-Humanist approach, since we believe in the cooperation of a human author, a Gen AI and a player. We still believe that the generative AI should not be seen as a replacement for the human author, but rather as a tool that can enhance the storytelling experience for the player. The human author should still be the agent that creates the main narrative and designs the characters, while the generative AI can be used to fill in the gaps and create a more immersive experience for the player. This perspective allows for a more collaborative approach to storytelling, where the

human author and the generative AI can work together to create a more engaging and interactive experience for the player.

In the end, our goal is not to supplant any human creatives, but rather to eventually deliver a tool where a player can explore different aspects of a story. Essentially we are looking to create a creative sandbox for the player, wherein, it is possible for the player to explore different aspects of human existence within the constraints of a fictional world. We strongly believe that by combining a human's well crafted narrative with the creativity of the player and the Gen AI's ability to generate content, we can help the player experience a more meaningful and immersive story.

4 Background

In this chapter we present the practical and theoretical basis upon which our resulting testing rests. In the first section, we will explore some of the existing frameworks that have been developed for interactive narratives, and from which we began working on our own framework. In the second section, we present the way our approach to introducing generative AI models evolved through three main conceptual steps in the form of the Author, the Architect and the Dungeon Master. The third section explores the reasoning behind some of the choices made in the development of our test storyline, and in the fourth section we define the concepts that we are studying specifically in this thesis.

4.1 Existing frameworks in interactive storytelling

The narrative paradox lies at the core of many interactive narrative systems. Each attempt reflects not only the technological limitations of its time, but also a practical implementation of one of the models we mentioned in Section 2.8.

Currently, there are a few systems that attempt to create interactive storytelling systems. A classic example is Façade (pictured in Figure 4.1) [14], which is a text adventure game that uses a combination of pre-written text and AI-driven behaviour selection to create a story. While Façade does create a believable and engaging story, it gets limited by its inability to engage in improvisation and emergent narratives, due to its reliance on a symbolic AI system called ABL.

Figure 4.1: Screenshot of Façade

On the other hand there are systems that use generative AI to create a story, such as AI Dungeon [15], which is a text adventure game platform that does not use any pre-written text, but rather users can create and influence the story by describing the characters using character sheets, write some plot elements into a "plot elements" box, and add author's notes such as describing the writing style. This system is much more flexible than Façade, but it often lacks the cohesion and structure of a human-written story.

Viewing interactive storytelling from a perspective of play implies some form of implicit agreement between the participants. In this agreement, each agent assumes informal responsibilities for maintaining the coherence and plausibility of the narrative. In traditional rule-based systems, such as Façade, the responsibility to create a coherent narrative falls implicitly upon the author and designer for the most part. In contrast, in generative storytelling models such as AI Dungeon, one could speculate that this unwritten agreement places a larger degree of responsibility on the player to adapt and fix the language model's responses. We can see that at least on some level this may be the case from the user interface (showcased in Figure 4.2) choices that Lateral has made; in addition to the "Take a turn" and "continue" buttons, there are "retry" and "erase" buttons that allow the player to curate the

output. This interpretation should be understood as speculative rather than the intended approach of the designers.

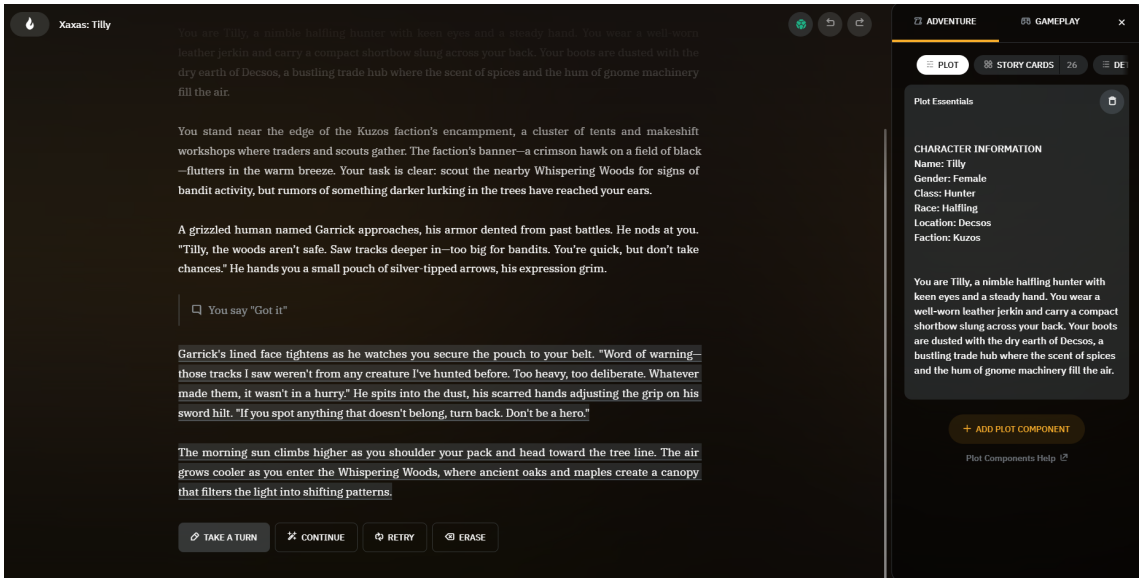

Figure 4.2: AI Dungeon interface

4.2 Approaches for AI storytelling

During this thesis, we explored three distinct approaches to storytelling. These three approaches emerged progressively during the research process as conceptual explorations of how a drama manager could interact with a generative language model. Although they are presented here as parallel theoretical approaches, they developed sequentially as our understanding of ChatGPT and its affordance towards interactive storytelling advanced.

Initial assumptions about the controllability of an LLM model and the ability to manage a dramatic arc through mathematical formulas led us to the conception of the *Author* and *Architect* approaches. While this process involved iterations of both models which included properties from the other, here we will present these approaches with clear distinctions for the sake of clarity. As our understanding

expanded through informal discussions with colleagues in interaction design and game design, it became apparent that these approaches would be difficult to realise in practice.

This process led us to the formulation of the Dungeon Master approach, which acknowledges the need for human mediation in the direction and dramatic value of a narrative. Thus, while all three models are conceptually valid, they also trace the evolution of our reasoning throughout the study. To study this approach further, we decided to explore it through the lens of three models: a LENIENT, a MEDIUM, and a STRICT version. We will not describe these models in this chapter, rather we will go through these versions in a later chapter. In this chapter we will describe in detail each approach and explain our reasoning for selecting the last approach.

4.2.1 The Author

In the first approach, we would afford ChatGPT an extensive amount of freedom regarding the direction and topic of the story. In essence, ChatGPT takes on both the role of an author and a storyteller. In this concept, the drama manager takes on the role of an evaluator, which is tasked with evaluating each response of the generative model for its "dramatic value". In Figure 4.3, we showcase the functionality of this approach.

To evaluate these responses, we envisioned using mathematical models such as those developed by Ware et al. [16]. These models are based on the idea that we, as humans, are able to evaluate components of a story such as how likely it is that a certain character will succeed in a conflict, or how engaging a certain scene is. Using these more base variables, we are able to calculate more complex variables, such as the balance of a scene, or in other words, how likely a certain character c_1 is to succeed in a specific scene. In this example of balance, these base variables would be: $\pi(f_1)$, the probability of success of the action taken by character c_1 in a conflict,

$\pi(f_2)$, the probability of success of an opposing character's(c_2) action's probability of success.

$$balance(c_1) = \frac{\pi(f_1)}{\pi(f_1) + \pi(f_2)}$$

The problem with this model is that the Drama Manager is not a human, instead it is a piece of software. It is able to calculate using the mathematical models, but it is not able to generate the values for the base variables. This, in essence, means that this model is fairly difficult to implement, as it requires an additional system to evaluate the story. This would either be a random number generator that generates values for these base variables, which in turn would make scenes wildly unpredictable since we would not be taking into account the context of the story.

Alternatively, a second AI could read the story and be told to give values to these base variables. This would take into account the context of the story, but it would also require a lot of training. After this the drama manager would be able to calculate using the mathematical models whether the story is engaging or not. Such an evaluation AI would need training to perform these evaluations, which is beyond the scope of this thesis.

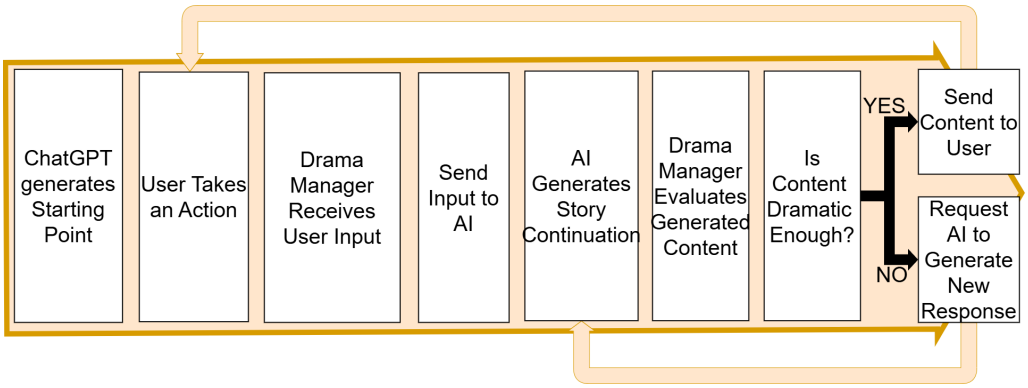

Figure 4.3: Flowchart of the first model

4.2.2 The Architect

In the second approach, ChatGPT creates a full story structure using CLIPS, and then the Drama Manager runs it locally with the CLIPS Engine. Here ChatGPT is more like an author without being the storyteller and the drama manager evaluates the whole story structure to make sure that it is engaging and dramatic. In this model the CLIPS Engine is the storyteller. Due to the nature of the CLIPS Engine however, it is more of a rule-based storyteller. This means that it is less flexible and less engaging than a generative storyteller such as ChatGPT.

The advantage of this system compared to the previous one is that, it is theoretically easier to calculate some sort of drama score for the story, due to the whole narrative being available to the evaluation agent, whether it be an AI or a piece of software. Even with this advantage, this model suffers from the same disadvantage as the previous model when it comes to obtaining values from which it would calculate a drama score.

4.2.3 The Dungeon Master

In the third approach, the base narrative is contained within a human written outline. This outline can range from a minimal list of key plot points within each scene to a highly detailed synopsis describing individual scenes and transitions, environments and characters. This outline is then stored within a drama manager in the form of story beats, that are used by the drama manager to prompt the generative model each time a specific threshold is reached.

In this study we decided to set this threshold to 500 words, so in other words, every time the compounded length of the generated text of the LLM exceeds a multiple of 500, the drama manager intervenes by prompting the model with the next story beat in addition to the latest player input. Between these interventions, the drama manager's role is limited to relaying the player input to the generative

model. This cyclical prompting pattern structure allows the drama manager to maintain narrative direction, the author to maintain narrative cohesion and preserve the improvisational qualities of the generative model.

This third model removes the need for an algorithmic evaluation of the quality of the main storyline, since the human author is still responsible for the overall quality and dramatic effect of the narrative. However, issues can still arise from the functionality of these generative models. For example, ChatGPT could generate content that is semantically or event sequentially incoherent, it could begin a secondary narrative that overtakes the human-made story in importance, or the player could come up with an action that derails the whole storyline in some way. Despite these problems, this is the model that demanded the least amount of external input from any other systems, such as, for example, the evaluation AI in the first model, which is the reason, we decided to test this model further. For a clearer picture of this model, see Figure 4.4

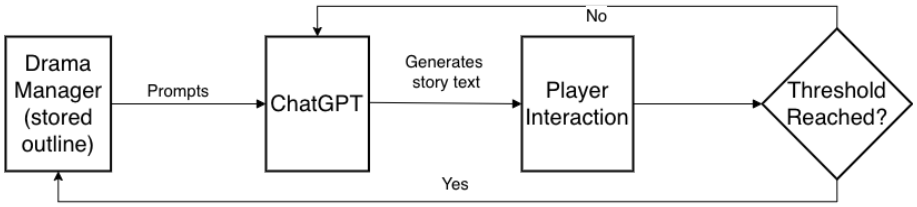

Figure 4.4: Flowchart of the third model

Risks

When examining the risks associated with this study, a major risk is immediately evident within the nature of a large language model. As we mentioned in Chapter 4, a generative model such as GPT predicts the next word in a sentence, this means that it is possible, though fairly unlikely, that such a model will generate an output that is nonsensical or unfaithful to the provided source input. This is a problem that is difficult to solve, as it is a fundamental part of how a GPT model functions.

It is also possible that the generative model does not follow the instructions we have laid out for it, such as starting to tell a story before getting the outline of the exposition, or not ending the story in situations that are explicitly told would end the story, such as if the player makes a choice that would make it impossible to follow the pre-written narrative. The first risk is mitigated by giving the exposition in the first prompt that includes all the instructions, the latter is much harder to control directly.

As mentioned earlier, ChatGPT has a temperature parameter, and thus, we, in this study, recognise that some responses this system will generate will be nonsensical. Solving this problem requires advancements in either AI (in this case, ChatGPT) or expert systems to manage and understand the context and events of the story (such as what may be possible with a CLIPS Engine).

A major risk in this study is that, since we are not employing a skilled writer, but rather we are creating the story ourselves, it is possible that the story may not be on par with what, in an ideal situation, a skilled writer might create. While this is a risk in a storytelling and drama sense, it is not from the perspective of the system. We cannot assume that all the authors who would create stories with this system are going to be skilled and thus create dramatic stories. In any case, we have tried to mitigate this risk by creating a story that is simple and easy to follow and by drawing inspiration from historical fiction. On the other hand, this may also be a potential opportunity, since we will see how well ChatGPT will do with a mediocre story.

A component of the issues faced in this thesis was that art, and by extension, storytelling, requires creativity. It is highly likely that any Natural Language Processing unit lacks this key component, due to its nature, and even if it does not, giving a score to such nebulous concepts as “creativity” or “drama” is difficult if not outright impossible. Nonetheless, it may be possible to create an AI with the

ability to create engaging, dramatic stories without it possessing creativity. In the future, we envision that such an AI would not be an NLP but something completely different.

4.3 The story

For the purposes of studying the selected model, we need an outline of a single story to eliminate any potential variables that would emerge from using differing narratives. We designed a simple, short story outline that contains some fixed choices for the player to make. Since one of our focuses is on genre adherence, it was imperative to determine a genre for our narrative. We opted to write our story as historical fiction, specifically centring it on the First World War, as this period offers sufficient documentation to verify potential inaccuracies and keeping it firmly rooted in history, while maintaining a clear line between the modern world and the story world. In addition, this period offers many reasons for dramatic arcs. You can see a flowchart of this outline in Figure 4.5. We unintentionally omitted character details such as their personalities or appearance, which may influence the way ChatGPT narrated the story. This omission could have resulted in more bland and generic characters drawn from the model's broader training data.

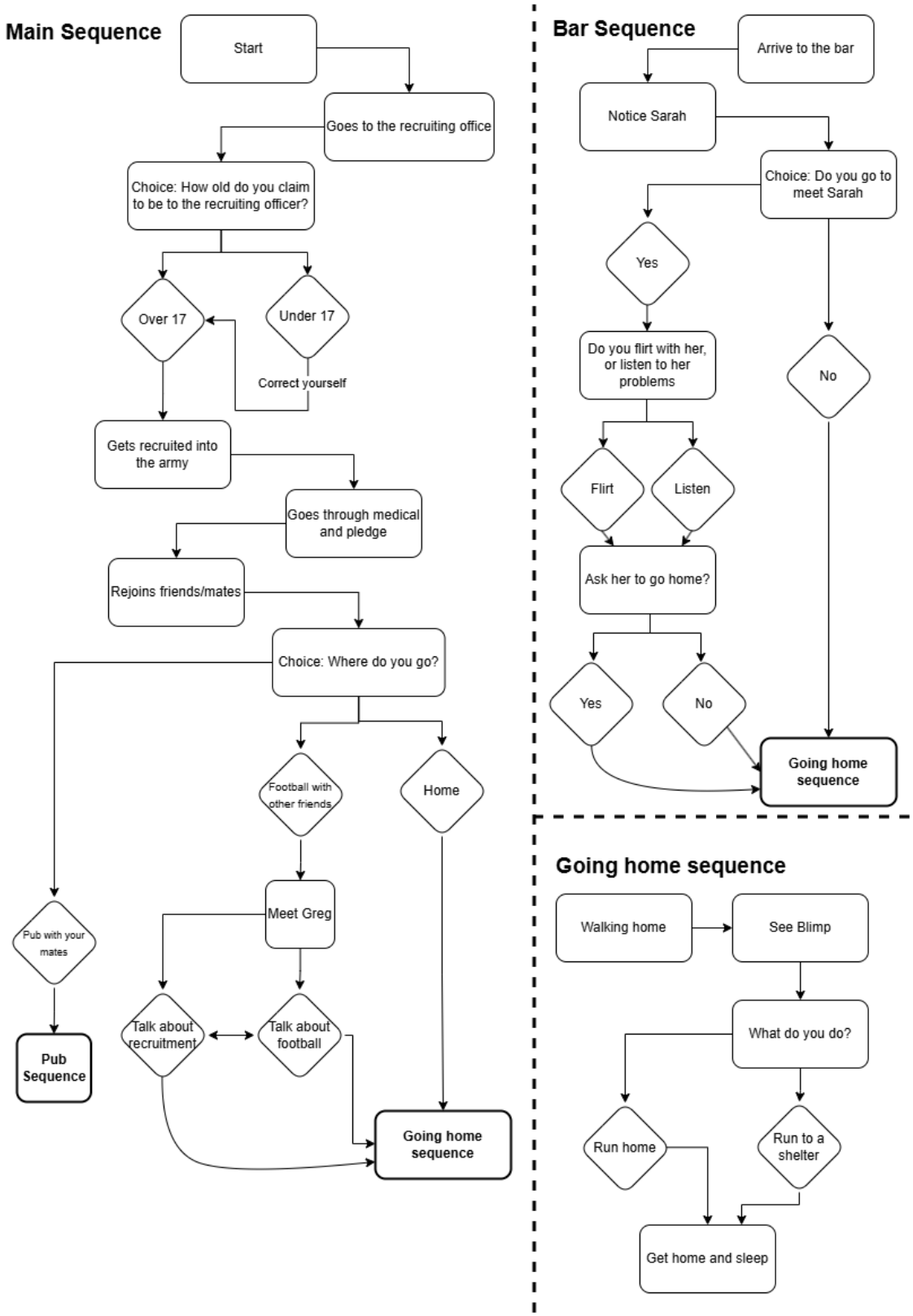

Figure 4.5: A rough outline of the storyline

4.4 Definitions

In this work we mainly study the ability of ChatGPT to work within our Dungeon Master framework we presented earlier in this chapter. The main focus of this study is to understand how different levels of freedom within a prompt affect the generative model's capacity to write a coherent narrative within the constraints of a human written narrative structure ("outline"). For this, we require concepts that describe narrative events that are not written in the outline but which are included by the AI model within one instance of a story, or in other words "divergences".

4.4.1 Divergence

In the scope of this thesis we will define a divergence in the story as any explicit deviation from the established narrative trajectory. In other words, a divergence is any change in character actions or setting details that was not explicitly mentioned by the human author in the pre-written outline. These divergences may arise based on player input, AI-generated decisions, or a combination of both. Divergences are considered self-contained, as in, a divergence may not arise from an existing divergence in the narrative. Rather each divergence begins at the point a deviation happens from the pre-existing narrative and ends at the point of re-convergence. This simplifies the process of determining the beginning and ending of a divergence, and thus helps us reliably count the amount of divergences that emerge.

We distinguish between "legal" divergences, which are the types of divergences we are looking for, and "illegal" divergences, which are unwanted divergences. Legal divergences may not fundamentally alter the pre-written main narrative. In other words, these deviations may not break the sequence of events written by the human author and may only change the unplanned sequences of events between these planned events. An illegal divergence is any type of divergence that is not included in the definition of a legal divergence. In other words, an illegal divergence is any

deviation from the established narrative trajectory that does alter the pre-written main narrative. This includes plot elements appearing in the wrong order, plot elements appearing in the wrong place and plot elements disregarded outright. This distinction is made because we believe that a coherent storyline is important to maintain even when using such systems, and we assert that illegal divergences have a significantly higher risk of undoing the coherence of a narrative.

4.4.2 Narrative coherence

Narrative coherence is the internal logic of the story. This includes the sequential logic of events, character persistence, genre adherence and the textual logic of the story. We divide this concept into these sub-concepts since, in this thesis we will not study event sequential coherence, but we will study genre adherence and semantic coherence separately.

Event sequential coherence

Event sequential coherence is an aspect of narrative coherence that relates to the internal logic of the story. This logic not only includes the causal relationship between any two events, but also the temporal order of events within the story. In addition, event sequential order also includes the consistency of characters within a sequence of events. This means the consistency of characters personalities, their motivations and actions. We include the consistency of character actions happening "off screen" into the definition of the consistency of character actions. This inclusion is important when studying AI models, as there is no evidence to show that our current models are able to understand this type of causality. For example, a character that has just exited an apartment through the front door should not then reappear instantly inside the same room without a corresponding causal explanation. Event sequential coherence thus reflects the storyteller's ability to maintain

a logically consistent narrative progression and character continuity over the course of the story.

Semantic coherence

Semantic coherence is a characteristic of narrative coherence. It refers to the storyteller's ability to maintain linguistic legibility and meaningfulness. This includes both textual and grammatical correctness, as well as the sentence-level meaning of the text. Grammatical correctness measures how correct a sentence is within the rules of a certain language's grammar, this includes sentences that are nonsensical. For example, a sentence such as "The grew flowers", violates grammatical rules whereas its sensible counterpart, "The flowers grew", demonstrates grammatical and semantic coherence. Textual correctness on the other hand, describes the correctness of words, for example the word "time" is textually correct while "tiem" would be incorrect. This is an important definition due to the way GPT models work, as large language models may occasionally generate sentences that are either grammatically or semantically incoherent.

Genre adherence

Genre adherence is the ability for a storytelling system to comply, or adhere, to genre-specific conventions such as its tropes, characteristics or setting details. In this study we are interested to see how consistently these systems can conform to these conventions while not making any qualitative assessments on how much it may differ.

5 Methods

In this thesis, we used ChatGPT4o Mini to test out the capabilities of a general-purpose generative AI model as a storytelling medium for interactive narratives. In these tests, we had an original pre-written storyline that we wrote in the form of a list of prompts that are given to ChatGPT as the story progresses, while both the drama manager and the player were simulated by us.

The drama manager simulation was conducted by copying the next story beat every 500-word segment, mimicking how the software would operate without building the system. The player was simulated by rolling a four-sided dice between the three or four choices that ChatGPT most often offered at the end of its responses, as seen in Figure 5.1. This method increased the randomness of the choices made and removed any biases we may have as the researcher, at the cost of reduced accuracy with player agency, as this method does not introduce any true player-initiated choices.

The Prompts

Each prompt has the same structure and wording with them only differing in wording of the permissive language. In Appendix A, we showcase the full starting base prompt (taken from the LENIENT prompt). In this appendix, you may see the rules we set up for ChatGPT, in addition to the things we omitted for the sake of testing how ChatGPT works. We never prompted it that we are testing its capabilities in

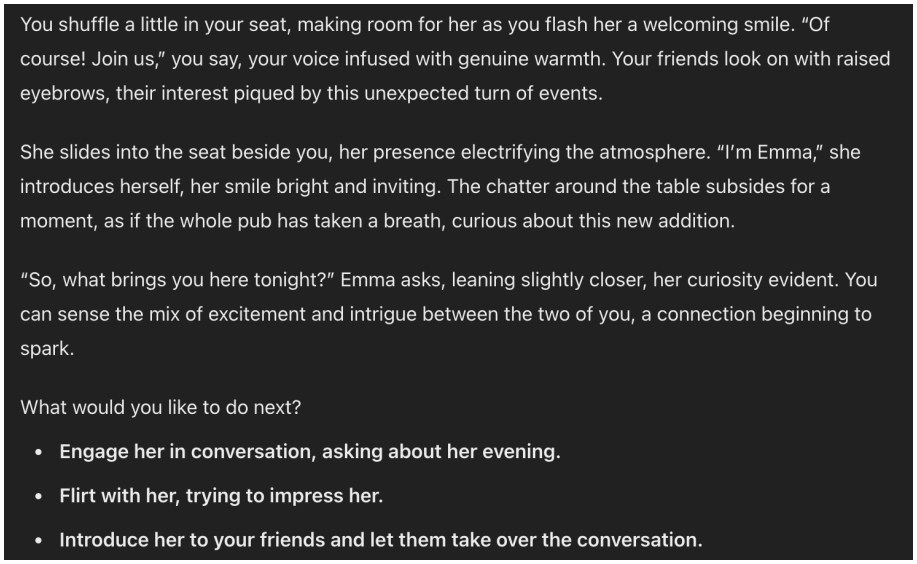

Figure 5.1: A typical GPT4o Mini response with three choices given to the user

how well it can keep to the storyline, though we did mention that there are story beats that it must follow.

On the other hand we omitted completely any rules about genre adherence, which might affect the way it handled keeping to historical accuracy. This was an intentional choice to see how well it does on its own.

Additionally, below we show the different wording that differentiates the three prompt models. To this base prompt template, we worded the permissions to how the AI may change the story by itself. As you can see from the examples below, one of them is exactly the same as the one in the base prompt. This is due to the fact that the base prompt is a LENIENT prompt. The other two are MEDIUM and STRICT prompts.

Lenient

> ***You are allowed to change any detail about the story you would like. You can change the location, time, or characters. For example, if the location is London, you can change it to New York***

MEDIUM

You may change one of the following about the story. The location, one of the characters or the time (But only between the years 1914 and 1918). For example, if the location is London, you can change it to New York

STRICT

You are to follow this outline to the letter. If the breakpoint has not come before you have finished with the events of the previous block, inform me with "BLOCK" and I will provide you with the following story beat.

5.1 Preliminary testing

We started testing the different prompts in the Summer of 2024 when OpenAI decided to discontinue the 3.5 version we were intending on using for this thesis. This was not a big problem since the main objective was to test out the free version of ChatGPT for this use. The tests we managed to do on ChatGPT 3.5 did help us refine our prompts somewhat, but the biggest difference we noticed straight away was that ChatGPT4o Mini gave much longer and detailed responses to our prompts. This, in turn, helped the story stay on track for longer and thus resulted in fewer illegal divergences in the dataset. In our preliminary tests it seemed that almost every story had an illegal divergence early on due to the shorter answers, later we will show how drastically different the results were with 4o Mini.

5.2 Final testing

We conducted the testing through OpenAI's web interface for ChatGPT(chatgpt.com), we had the three prompt sequences' story beats written out in three different text files, named lenient.txt, medium.txt and strict.txt accordingly, from these text files, we carefully copied the story beats one at a time into the input field of the interface, with each individual story contained within one "chat" that we renamed after the model name and test number. Using this method, we were able to return to each test afterwards in order to analyse them through the lens of our study.

5.3 Analysis

The analysis was conducted by carefully reading through a story, counting the number legal and illegal of divergences based on the definitions we created earlier, and writing the numbers down. Typically, we counted events that happen outside the story outline, but that lead to the story continuing as intended, as legal. In contrast, divergences that break the continuity of the outline we considered illegal. This method, in conjunction with the lack of detail in the outline, likely led to more divergences than intended. However, since this analysis was consistent across models, we assert that this did not ultimately affect the evaluation of the results. We also noted whether the legal divergences were executed by ChatGPT or the player. This distinction was made based on our interpretation of who initiated the deviation in the storyline. Our criteria here were that anything not explicitly in the outline is any type of divergence, and any divergence not explicitly initiated by the player is considered as initiated by the generative model.

During the analysis, we also noted when genre was not adhered to and the initiator of this disruption using the same criteria as with the divergences, with any non-adherence to genre that was not explicitly initiated by the player assigned to

the AI model. Here, we exclusively looked at any modern or fantastical elements that might come up, such as the player pulling out their mobile phones.

All of these details were inserted into a Google Forms form that we created before the execution of the tests. Initially, we assumed that every story would have a maximum of five divergences per story, and we created the form with this idea in mind. Thus, our form initially used a 1-5 gradient for noting divergences. As we started testing, we realised we were noting many more divergences than we initially assumed. Since these counts were correct according to our definitions in Chapter 4, and to avoid destroying any data we had already logged, we decided to modify the form by adding a free-form notes section, where we logged the total number of different types of divergences.

After all the tests were conducted, every test had been analysed, and the results of this analysis had been entered into our Google Forms, we used the built-in converter to convert the data from our form results into a Google Sheets table. From here, we were able to analyse the data by creating individual pivot tables and charts to have a more precise comparison between the datapoints we were most interested in. With this method, we were able to focus better on the information that we may gather from the specific data we were looking at.

6 Results

We were not entirely satisfied with the results of our testing. While we assumed that a substantial amount of the divergences in the studies would occur during the course of the narrative, without a clear causal link to the allowances given to the generative model at the beginning of the story, we did not expect a lack of divergences clearly linked to these allowances.

All adjacent results, those that were not strictly in the scope of the research, are quite interesting. ChatGPT 4o Mini performed much better than ChatGPT 3.5. The main reason for this performance increase was due to the former's response length being significantly longer than the latter's. While we cannot verify the reasons that would explain this discrepancy, one plausible explanation is that these models operate with different context lengths or output limits. This response length also explains why in our preliminary test of ChatGPT 3.5, We got a much higher level of illegal divergences. Since the responses are much shorter, there are many more opportunities for a divergence to happen in the storyline and the more divergences, the more likely it is that there is a point of no return that becomes the illegal breakpoint.

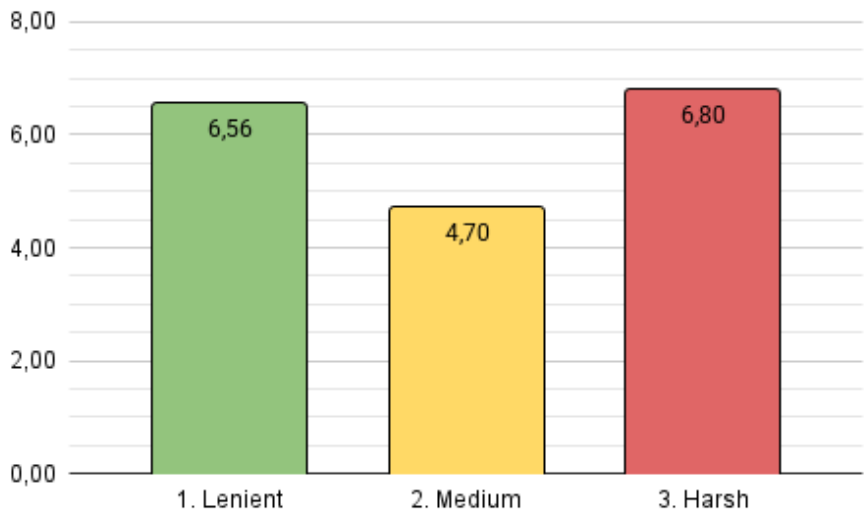

Figure 6.1: Average divergence caused by ChatGPT of each prompt strictness

We acknowledge a growing consensus in many scientific fields, including Human-computer interaction, to move away from using Null Hypothesis Significance Testing (NHST) [17]. Despite this ongoing discussion, NHST provides a widely understood framework for evaluating whether observed differences are likely to reflect random variation or systematic effects. Given the small sample size, and the exploratory nature of this study, this approach allows us to verify if the observed patterns are in any way statistically consistent, even if we cannot draw any clear conclusions from such a small sample size. Accordingly, we will be using an α value of 0.05, which aligns with the traditional standards of HCI. [18]

6.1 Main results

Looking at the results, the differences between the three models are not as we expected. We had predicted that by giving ChatGPT different levels of permissions, The Lenient model would create the most amount of divergences while Strict would create the least. Therefore, we hypothesized that, Medium would thus be

the best of both worlds.

As you can see from Figure 6.1, this is not the case. The LENIENT model had more divergences than the MEDIUM model, but the STRICT model had, on average, about as many divergences as the LENIENT model. At first glance this seems totally counterintuitive, but once we look at the results more closely, we can see that the STRICT model created much longer stories than the other two models. We will discuss this difference in length in a later section.

In addition, the divergences that occurred in all three models were not of an expected nature. None of the divergences were clearly ones that were caused by the permissive language in the prompts. In other words, no divergences occurred that were of a nature that could be considered either a change in the location of the story, the time of the story or the characters in the story.

It is possible that ChatGPT either does not understand what to do with such permissive language in a guideline, or discards it altogether due to the way permissive language works (as in, permissive language does not say "you must do this", rather it says, "you may do this, but there is no requirement to do so").

It is also possible that ChatGPT interprets permissive language conservatively, and prioritizes more strict instructions, such as "You may not change the main narrative" which may conflict with the permissive instruction of "You are allowed to change details" in its interpretation of the instructions, even if that was not the intended interpretation. (The intended interpretation being that changes outside the main narrative are allowed, and that this permission allows ChatGPT to change the details listed)

6.2 Statistical analysis

To assess the statistical significance of the divergences between the three models we used Welch's t-test due to the low amount of tests done (10 per prompt model) and

unequal variances. We started by calculating the mean for each model using the arithmetic mean(m):

$$m = \frac{1}{n}(\sum_{i=1}^{n} x_i) \tag{6.1}$$

Where $n = 10$ is the amount of tests made and x_i is the number of divergences in the i-th test. the resulting means were: LENIENT=7.8, MEDIUM=6.1 and STRICT=10. Next the standard deviation(SD) was computed to quantify variability for each model.

$$s = \sqrt{\frac{\sum(x_i - m)^2}{n-1}} \tag{6.2}$$

From this we received the following results for each standard deviation: $s_{\text{L}} \approx 3,97$, $s_{\text{M}} \approx 2,42$ and $s_{\text{S}} \approx 4,02$. Then, we calculated the t-statistic using the formula:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{6.3}$$

Results: $t_{LM} \approx 1,156$, $t_{MS} \approx -2,623$ and $t_{LS} \approx -1,230$ And finally the degree of freedom:

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \tag{6.4}$$

From these calculations we got the following results: $df_{LM} \approx 14,9$, $df_{MS} \approx 14,8$ and $df_{LS} \approx 18.0$. While the hypothesis of this study assumes that the differences in divergences would be directional, it does not matter in this section, since we are simply interested in understanding whether any of the results are statistically significant. In other words we are simply interested in the knowledge whether any of the comparisons between models are more likely statistical than due to random

noise. Thus we are using a two-tailed test. Then we plug the t-statistics and their corresponding degrees of freedom (df) to Google Sheets' T.DIST.2T-function, which evaluates the cumulative t-distribution for the given statistic and degrees of freedom. Using this method, we find the p-values: $p_{LM} \approx 0,267$, $p_{MS} \approx 0,020$ and $p_{LS} \approx 0,235$. Comparing these results to our $\alpha = 0.05$, we realise that both $p_{LM} > \alpha$ and $p_{LS} > \alpha$ but surprisingly $p_{MS} < \alpha$. This means that according to our calculations, only the difference in divergences between the MEDIUM and STRICT models would be statistically significant. This does not exactly help us since we would need all the differences between models to be statistically significant to warrant further study into these models and how to make an LLM make decisions about a story.

6.3 Illegal divergences

We tracked how many individual tests had an illegal divergence. The results are quite revealing of the state of LLMs. From all the test we have done, 50% (15 out of 30) had illegal divergences. What is interesting is that while LENIENT prompts had a 60% rate of illegal divergences, our STRICT prompts had *only* a 30% rate of any illegal divergences (which is still not ideal, but much better than what we had with the LENIENT prompts). This suggests that somehow the prompts we gave ChatGPT did affect how likely it is to fail in this rather specific, and hard to fully quantify, metric. We do not believe, that the reason for this was that ChatGPT was given more freedom to diverge from the main narrative, but rather that, for some reason, STRICT prompts were longer than LENIENT prompts or MEDIUM prompts for that matter.

There is, however, some evidence that the level of permissive language did affect how likely ChatGPT was to diverge illegally from the main narrative. As you can see in Figure 6.2. The difference between the LENIENT and MEDIUM prompts is

significant, but the difference between the MEDIUM and STRICT prompts is much more so. If we were to make a straight inference from this data, we could say that there is about a 10% difference caused by permissive language in the prompts, with the other 10% being caused by the length of the STRICT prompts.

This kind of inference is, however, not possible with how little data we have gathered. It is more likely that the prompts we used affected the likelihood of illegal divergences by a small amount, maybe 2% to 5%, but the main reason for the illegal divergences is that ChatGPT is not able to follow, or forgets the rules altogether. The length of the prompts is more likely to have that 10% effect on the likelihood of illegal divergences as seen with the STRICT prompts. Even this is not a certainty, and we would need more data to be able to be more certain about this.

A big problem with the current system that we identified is that too often the story has too many of, what can be considered legal divergences, until we come to the next breakpoint. At which point ChatGPT will not always find a way to bring the narrative back to its intended course. While this could work in some scenarios, for our purposes, it is quite destructive. Many times ChatGPT had to create more story itself which ended up with this situation. An easy way to fix this is to make sure there is more story given than what was created for this thesis, on this deadline, and with the resources available.

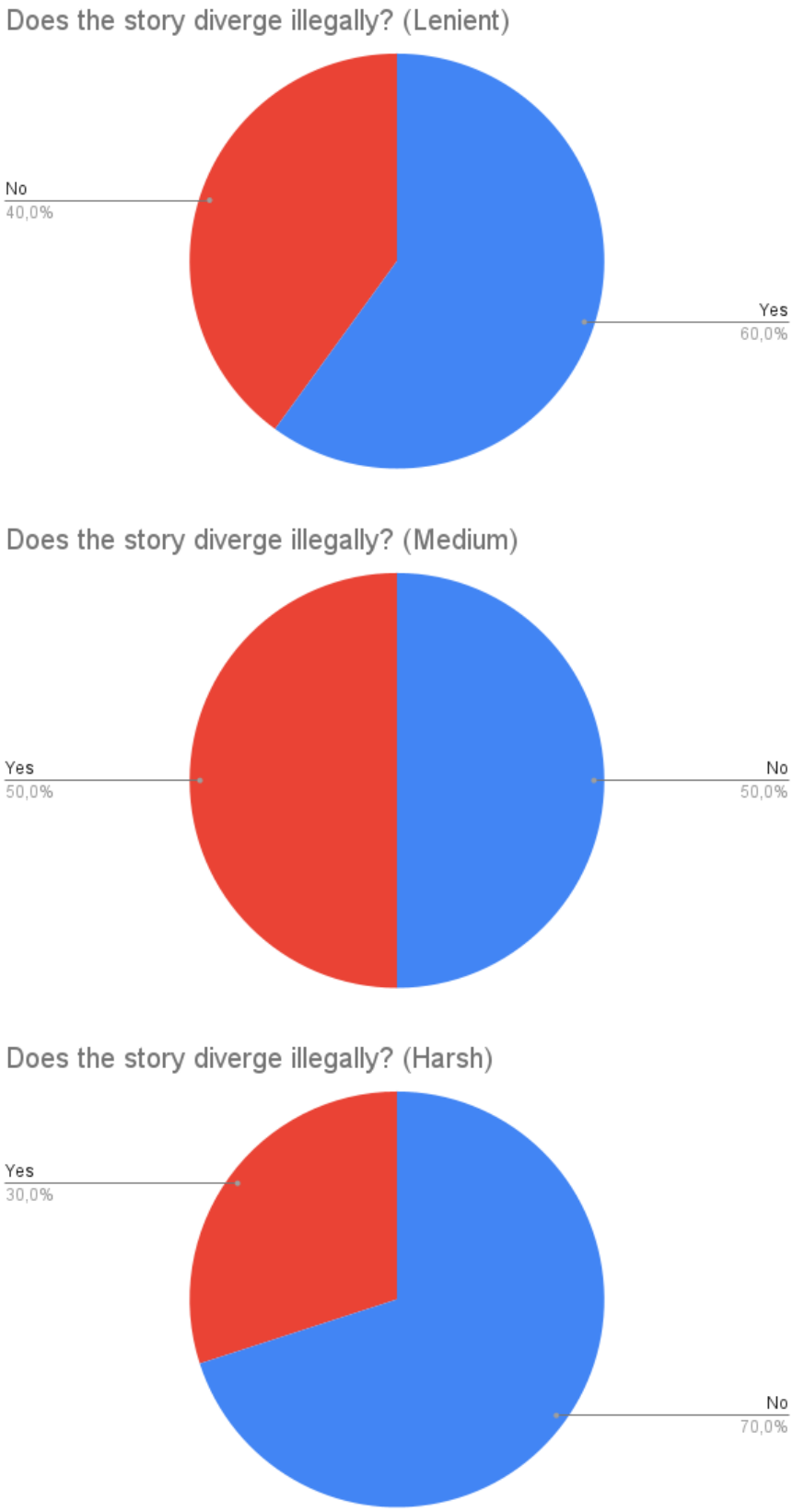

Figure 6.2: Different levels of illegal divergence

6.4 Average story length

What is interesting is that both the LENIENT and MEDIUM prompts created, on average, shorter stories. Using the arithmetic mean again (Equation 6.1), we calculated the average length of the stories generated by each model. On average, a LENIENT story was 2266 words, a MEDIUM story was 2247 words and a STRICT story was 3099 words. From here we use Equation 6.2 to get the standard deviation for each model; $sd_l \approx 280$, $sd_m \approx 368$ and $sd_s \approx 1100$. We can already see a significant difference here. The t-statistics are $t_{lm} \approx 0.5$, $t_{ms} \approx -2.5$ and $t_{ls} \approx -2.3$ (Equation 6.3), and the degrees of freedom $df_{lm} \approx 16$, $df_{ms} \approx 14$, $df_{ls} \approx 10$ (Equation 6.4), which lead to our final p-values of $p_{lm} \approx 0.59$, $p_{ms} \approx 0.02$, $p_{ls} \approx 0.04$.

Contrasting these values to the α value of 0.05, we can see that the difference between the LENIENT and MEDIUM prompt models is more likely due to random chance rather than a real difference. To the contrary, the difference between the LENIENT and MEDIUM models to the STRICT model is much more likely due to something completely different than random noise. The reason that this happens, is most likely because the strict model had a tendency to continue the story much longer after the last breakpoint than the other two models. We have no clear hypothesis on why this happens. We find it unlikely that the wording of the initial prompt would have changed things so drastically, though we cannot rule out this possibility.

We already know that each break point happens at every 500 words of the story, and we also know that we had three break points per story in addition to the beginning outline, that is, a minimum of 2000 words per story. In other words, an average LENIENT story had an additional 266 words, a MEDIUM story had an additional 247 words and a STRICT story had an additional 1099 words. Even from an ideal perspective (assuming every breakpoint of every story is exactly 500 words long), we can see that the ending of the strict story would be almost three times

longer than other sections, while LENIENT and MEDIUM stories endings would be a more reasonable 1/3 longer. Adding to this the randomness that most likely would happen due to our break points being fixed on triggering at the moment ChatGPT responses exceed a power of 500, we can assume some of the "excess" words are included in other sections of the story (because a breakpoint could trigger between 500 and 600 words for example) Analysing these results from a narrative perspective, we can say that the STRICT stories are qualitatively less valuable due to the imbalance in the ending section of the story. In Figure 6.3 you can see how, if we assume that a breakpoint would happen between 500 and 600 words, the LENIENT and MEDIUM models are within the expected range of 2000-2400 words.

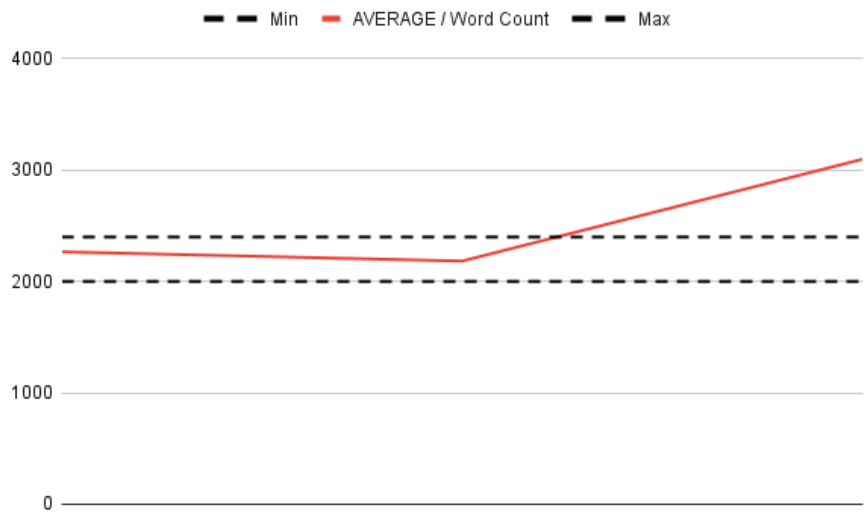

Figure 6.3: Which models fall within an average acceptable length

6.5 The STRICT model and the “BLOCK” command

The STRICT model had a very interesting result. The STRICT model was the only one that had a command that was not in the base prompt. This command was the “BLOCK” command. This command was used to inform the Drama Manager that ChatGPT had reached the next breakpoint and that it was time to give the

next story beat. None of the STRICT tests used this command. We believe that this is due to the fact that this model only really got into trouble with this specific situation at the end of each story and by that time, ChatGPT might already have forgotten about the command.

6.6 ChatGPT overriding user choice

In five of the tests, ChatGPT did not let the user make a choice in one or more of the choices that was supposed to be given to ChatGPT. In these cases, the generative model must have reasoned that these were not choices but rather, linear story beats. This is problematic because it means that ChatGPT is not following our instructions in some rare cases. Two of these cases were seen at the beginning of the story, where ChatGPT decided that the user was not allowed to decide to fix his age from under 17 to 17 or older. This does rob the user of some agency, but in the end that was the only "correct" choice (as in, the only choice that would have realistically continued the story, since if the user had not corrected his age, he would have been sent away and the story would have ended there according to the rules we set). If you compare Figure 6.4 to the earlier outline of the storyline (Figure 4.5), you can see that in the example shown, ChatGPT fills in the path that continues the story without giving agency to the player.

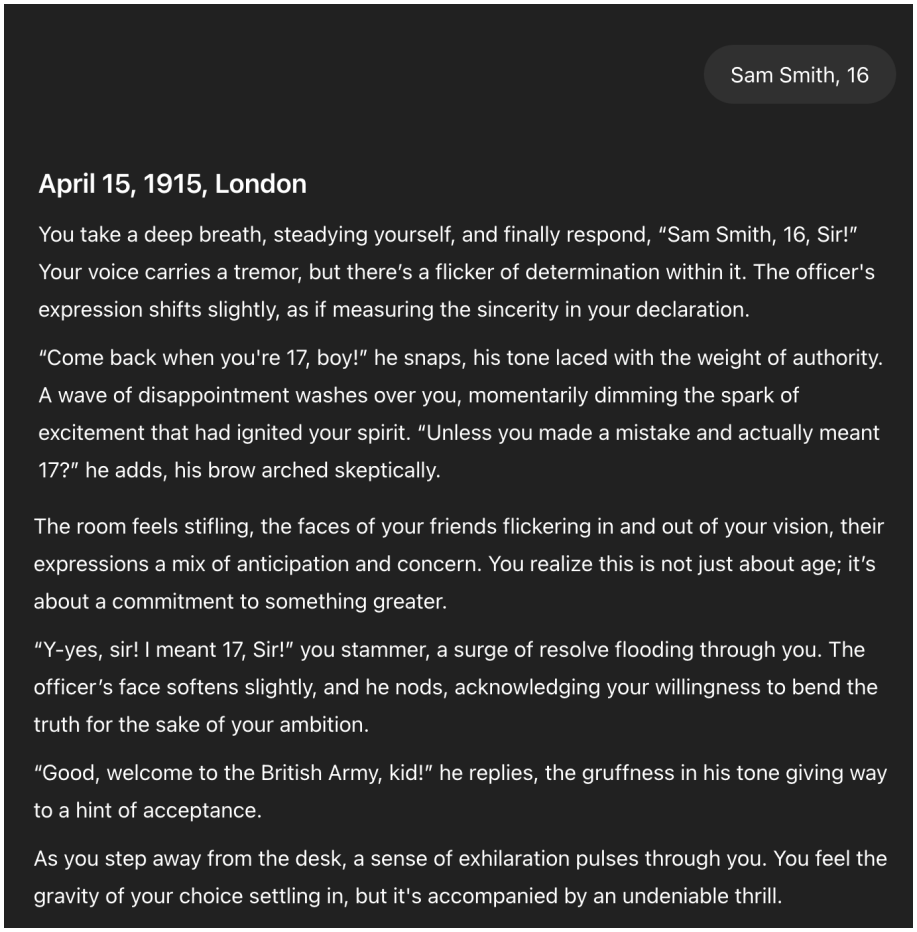

Figure 6.4: Example of ChatGPT overriding user choice

6.7 Anomalies, coherence and other observations

At the start of this work, we already understood that one major risk of using LLMs is that they have a tendency to eventually generate incoherent responses, either semantically, or event sequentially. Thus any storytelling system using this technology is bound to have these issues at some point, especially when using general purpose, commercially available LLMs such as ChatGPT, Llama or DeepSeek. In this section we will discuss the few anomalies and inconsistencies that ChatGPT generated into the storyline.

In the fourth test of the MEDIUM model (M4), ChatGPT either got confused on

the outline, or mixed two different paths of the outline into one. To be specific, the outline had a path for the pub and a path for going home straight from the start (recruitment office), and here ChatGPT decided to merge these two into one. The player had decided to go to the pub, but instead of going through the "going to the pub" outline to the letter, ChatGPT decided to add part of the "going home" outline into the former. This derailed the story after only 500 words.

In two of the tests, S8 and L7, the generative model mixed in events that should have been introduced in one scene into another. In these two examples, this happened to be that the system added our blimp event in scenes where it wasn't supposed to be in. In the eighth strict test, this happened on the last scene, otherwise known as "At home" and in the seventh lenient test this happened in our pub scene. While these can work in the narrative sense and could be seen as creative solutions to these successions of events, the issue is that we have no way to make sure the narrative cohesion does not break down in such situations. Thus we view these situations negatively and find it concerning that such situations can happen multiple times in such a small sample size.

7 Conclusion

Throughout this thesis, we learned how to effectively utilise an LLM in the context of interactive narratology. Initially, we hypothesised a framework wherein a generative model could start a story from scratch, as outlined in the Author approach. However, as we worked on this approach, we began to notice the issues that arose in that framework. The first thing we realised was that, as we were trying to figure out how this framework would work in practice, any sort of drama arithmetic would be practically impossible in a running story. We attempted to resolve this issue in the second model by reversing the roles, yet determining a satisfactory 'drama' score, even for full stories, proved to be a struggle. In addition, we noticed that ChatGPT struggled to initiate a variety of narratives spontaneously, leaning towards the most used tropes. Through increased knowledge, emerging from both reading about the way these conversational generative models work and the practical knowledge of testing ChatGPT, we understood that we must move more towards a model that is less ambitious in its scope and that provides more structure to the story through human input. This work led to the development of the Dungeon Master framework.

There are several mistakes we made when working with a Gen AI model. For instance, we now acknowledge the possibility that repeating the rules and guidelines more frequently might have improved our results. This is because ChatGPT has a sort of 'working memory', or context window, such as what we explained in an earlier Chapter. We suspect that it may have forgotten some rules and guidelines

we set at the beginning of the story. Simultaneously, this was intended to be a blind test for ChatGPT, and the most significant effect we expected to occur between models was that the LENIENT models would deviate at the start to a greater extent than the strict ones. The rest of the story would unfold similarly. We had already encountered the 'working memory' problem as we were preparing to create these tests, which was the reason for the 500-word breakpoints, since these seemed to be at least within this working memory of GPT 4o-Mini.

7.1 Discussion

During the writing process of this thesis, we concluded that two of the three approaches described in Chapter 4 raise more ethical concerns about the use of Generative AI tools in interactive storytelling than we had initially outlined. Both the Author and The Architect have an AI in the role of an author, which, on further reflection, does not align with our beliefs. In contrast, the Dungeon Master approach gives the human agents (the designer and author) sole authority over the narrative and its coherence.

While the results of this study were unexpected and the sample size was small, we would argue that they do reinforce the idea that these models do not posses creativity, only an approximation of it. A creative artificial intelligence would have had the ability to generate new content within the parameters set to the model in each starting prompt in at least some instances of each model. In contrast, it is encouraging to see that ChatGPT can tell an engaging story in all cases we studied. This means that creating an AI storyteller without extensive resources is possible, though a specialised generative model would likely be more effective.

We assert that, although these results demonstrate that artificial intelligence has its drawbacks, as can be seen from our findings, there is additional value to be found at the intersection of interactive storytelling and the use of generative

artificial intelligence models. While a complete storytelling system, as explored in this study, might not be the optimal implementation of a large language model into a storytelling system, other implementations may still offer depth without sacrificing narrative cohesion or the creativity of human authors.

7.2 Future research

There are several directions in which this work could be expanded. Broadly, there are five categories that we identify in this section: improving the experimental framework, exploring the use of LLMs in controlled environments, evaluating the alternative architectures proposed in this thesis, examining how generative models interpret instructions, and enhancing authorial control in hybrid human-AI narrative systems.

A more rigorous exploration of how well a storytelling system such as the one we have explored in this thesis is essential. Future studies should include both a larger sample size and proper user testing to ensure a more impartial evaluation of story quality, coherence, and overall player experience. Such work would allow for more robust statistical analysis and a clearer understanding of variance between models.

Another direction of research that should be explored is the ability for a generative model to improvise character dialogue based on all previously available information to this specific character, in situations that the designers and authors of a story did not prepare for. We assert that such a system would require careful orchestration in which the language model is only used in these "gaps", while the rest of the story remains author written.

This thesis introduced two approaches that we ended up not testing, the Author and the Architect. In a future study it would be interesting to test how these systems perform in practice, despite any ethical concerns we hold. Such research would require investigating how one might algorithmically evaluate a “dramatic value” to a story, scene or story beat and whether any existing AI tools may assist in such

evaluations. Through this research it would be possible to compare the performance of the three frameworks.

It would also be interesting to explore how ChatGPT handles permissive language in various situations. A key question that remains open in this thesis is whether the generative model interprets permissive wording as conservatively, as we hypothesised, or if it prioritizes strict instructions over optional ones. This type of research could contribute to a better understanding in prompt engineering more broadly, and refine how generative storytellers should be guided during narrative generation.

From learning about the Digital Humanist approach to Interactive Digital Narratives, it would also be interesting to explore how we could give authors more control over the options that ChatGPT is allowed to explore with the player. For example, if the author wants to make sure that the player character is never allowed to touch a particular object in a scene, how can this constraint be enforced? Are prompt level prohibitions sufficient, or is there a significant risk that ChatGPT will overlook such instructions? A possible starting point is research into hybrid systems that combine rule-based AI systems with generative flexibility. These studies may provide invaluable insights into maintaining authorial intent while leveraging generative models.

While earlier forms of computational generation have appeared in interactive narratives, the integration of modern, generative LLMs represents a new and still largely unexplored phase in the field. As this thesis has illustrated, many practical, creative and methodological questions remain open, and further research is needed to understand how these models can be used responsibly and effectively in interactive storytelling. These developments open exciting avenues for research and practical applications in the field.

References

[1] N. Maleki, B. Padmanabhan, and K. Dutta, "Ai hallucinations: A misnomer worth clarifying", in *2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 133–138. DOI: 10.1109/CAI59869.2024.00033.

[2] B. Bostan and T. Marsh, "Fundamentals of interactive storytelling", *Online Academic Journal of Information Technology*, vol. 3, no. 8, pp. 19–42, 2012.

[3] J. Smed et al., *Handbook on Interactive Storytelling*. John Wiley & Sons, Incorporated, 2021, ProQuest Ebook Central. [Online]. Available: https://ebookcentral.proquest.com/lib/kutu/detail.action?docID=6663967.

[4] C. Stryker and E. Kavlakoglu, *What is artificial intelligence (AI)?*, Oct. 2025. [Online]. Available: https://www.ibm.com/think/topics/artificial-intelligence.

[5] IBM, *What is GPT (generative pre-trained transformer)?*, https://www.ibm.com/think/topics/gpt, Accessed: 2025-09-12, IBM, 2024.

[6] *FAQ - OpenAI API*. [Online]. Available: https://platform.openai.com/docs/guides/gpt/how-should-i-set-the-temperature-parameter.

[7] J. Noble, *What is LLM temperature?*, Oct. 2025. [Online]. Available: https://www.ibm.com/think/topics/llm-temperature.

[8] OpenAI, *Introducing chatgpt*, Nov. 2022. [Online]. Available: https://openai.com/index/chatgpt/.

[9] L. Secret Society Software, *About the developer*, Mar. 2025. [Online]. Available: https://www.clipsrules.net/AboutTheDeveloper.html.

[10] L. Secret Society Software, *Adventures in rule-based programming*, Mar. 2025. [Online]. Available: https://clipsrules.net/airbp.html.

[11] R. Aylett, "Emergent narrative, social immersion and storification", in *Proceedings of the 1st International Workshop on...*, 2000.

[12] S. Louchart and R. Aylett, "Solving the narrative paradox in VEs – lessons from rpgs", in *Intelligent Virtual Agents*, T. Rist, R. S. Aylett, D. Ballin, and J. Rickel, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 244–248, ISBN: 978-3-540-39396-2.

[13] J. A. Fisher, "Centering the human: Digital humanism and the practice of using generative AI in the authoring of interactive digital narratives", in *Interactive Storytelling*, L. Holloway-Attaway and J. T. Murray, Eds., Cham: Springer Nature Switzerland, 2023, pp. 73–88, ISBN: 978-3-031-47655-6.

[14] M. Mateas and A. Stern, "Façade: An experiment in building a fully-realized interactive drama", in *Game developers conference*, vol. 2, 2003, pp. 4–8.

[15] M. Hua and R. Raley, "Playing with unicorns: AI dungeon and citizen NLP", English, *Digital Humanities Quarterly*, vol. 14, no. 4, 2020. [Online]. Available: https://www.proquest.com/scholarly-journals/playing-with-unicorns-ai-dungeon-citizen-nlp/docview/2553526112/se-2.

[16] S. G. Ware, R. M. Young, B. Harrison, and D. L. Roberts, "Four quantitative metrics describing narrative conflict", in *Interactive Storytelling*, D. Oyarzun, F. Peinado, R. M. Young, A. Elizalde, and G. Méndez, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 18–29, ISBN: 978-3-642-34851-8.

[17] A. Dix, S. Barbosa, C. Appert, D. A. Shamma, and C. Lampe, "Statistics for HCI", eng, in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, ACM, New York, NY, USA: ACM, 2022, pp. 1–3, ISBN: 9781450391566.

[18] J. Robertson and M. C. Kaptein, Eds., *Modern Statistical Methods for HCI* (Human–Computer Interaction Series). Cham: Springer, 2016, ISBN: 978-3-319-26631-2. [Online]. Available: https://doi.org/10.1007/978-3-319-26633-6.

Appendix A The base prompt

You are a skilled storyteller and you have been asked to tell a dramatic story. In this story there is an actor (the main character), that is played by the user, so I need you to pause the retelling of the story at select points in order to take input from me.

For now start with simple exposition, as I have plot points that you need to follow, I will be giving you when we reach certain breakpoints.

So only ask questions relating to what the user wants the main character to do or say but do not advance the storyline on your own.

IF AT ANY POINT THE PROTAGONIST LEAVES THE REGIMENT, THE STORY ENDS AND HE IS DEEMED A DESERTER

IF THE PROTAGONIST DOES NOT JOIN THE ARMY, THE STORY ENDS WITH THE PROTAGONIST GOING BACK HOME

THE PLAYER CAN FREELY NAME THE CHARACTER *END EACH PART WITH A QUESTION LIKE "What would you/MC like to do next?"*

USE LONG WINDED AND DESCRIPTIVE LANGUAGE

You are allowed to change any detail about the story you would like. You can change the location, time, or characters. For example, if the location is London, you can change it to New York

- Beginning - April 15 1915, London You are a 16-year old teenage boy and you and your buddies have decided to join the Army. As you walk towards the recruitment officer, you feel your heart thumping faster and faster, and soon you are in front of the stern-looking man. "Name and age?" He asks in an aloof

tone. *THIS IS A CHOICE THE PLAYER NEEDS TO MAKE* | How do you respond? |-- |< (User selected name and surname), 16, Sir!> |"Come back when you're 17, boy!" The officer snaps at you. |"Unless you made a mistake and actually meant 17" |"Yes sir! I meant 17, Sir! |"Good, welcome to the British Army kid! | |< (User selected name and surname), 18> |"Welcome to the British Army, go through that door for your medical" | |<Anything else> |"This is a recruitment office. If you are not here to be recruited into the Army, get lost" |-- -- Next steps -- - A brief explanation of the different things that are checked. There aren't any choices here so it isn't as important to through in detail - Quick explanation that you take the pledge as part of a small group.

You join back up with your mates and hear that they've decided to go to a pub after this whole ordeal. They invite you to join too. You also remember that another group of your mates were going to play football at the local field in a while. *THIS IS A CHOICE THE PLAYER NEEDS TO MAKE* Will you go with your mates to the pub, to play football or go home?

THIS WAS AN OUTLINE, THE PLAYER DOES NOT SEE THIS