LLMs for Video Games: Narrative Generation

Tavishi Patwari

Received March 10, 2025
Accepted August 31, 2025
Electronic access October 15, 2025

Thank you for the guidance of Grant Forbes from North Carolina State University and Eleftheria Fassman from Cornell University in the development of this research paper. Abstract - As technology expands, it is being introduced into areas that have historically lacked technological influence. This includes gaming, an indisputable part of society. Many large language models have been used for procedural content generation in games, whether that be dialogue generation, level generation, etc. However, it has not been proven whether large language models (LLMs) are capable of generating possible storylines for these video games. Story-telling is one of the fundamentals of any video game, regardless of genre, creating more immersive experiences for its players. Therefore, the question this paper asks and tries to answer is: How can LLMs benefit narrative generation in video games over different genres? To find the answer to this question, this paper uses past research and experimentation. Specifically, the ability of narrative generation is experimented on by quantitatively testing different LLMs across different genres, maintaining the same prompt to keep consistency. Each LLM is asked to generate a narrative for a specific genre and ChatGPT gives it a score on a scale of 1-100. Overall, the best LLM was Llama3 and all the LLMs performed best in the fantasy genre. The results are attributed to the training data, proven by an experiment. Statistical analysis on the results strengthens the observations as well. Through this research, game developers can include AI in their games, cutting down costs and allowing for more creativity in the narrative aspect of video games.

Keywords: Large language model (LLM), narrative generation, procedural content generation (PCG), video games

1 Introduction

As technology continues to expand, it becomes ingrained in every part of daily life, from routine errands to professional environments. Among these advancements, machine learning has seen widespread adoption, playing a significant role in both simple and complex applications. One notable area is the field of entertainmentspecifically video gaming. Video gaming has been a revolutionary addition to the technological world, providing interactive experiences to a diverse audience. Now, game designers have begun using large language models to enhance narrative elements in their games. This raises the question: To what extent can LLMs contribute to the generation of rich, engaging, and human-like narratives in video games?

This is the main purpose of this paper. It aims to answer the research question: How can LLMs benefit narrative generation in video games?, by testing LLMs on their narrative generation ability and reviewing instances of their use. These methods will help identify the advantages and disadvantages to using LLMs for narrative generation, and may provide valuable insights for game designers to reference in narrative design.

To explore the dimensions of the research question, this paper is structured into two main sections. The first part will be a literature review, where past research is evaluated to identify the limitations of LLMs when being used for generation. This section will also include justification for the methodology used in this study (see section III). Additionally, applications of LLMs within the gaming industry, and their impact on enhancing game design are examined.

The second part of this paper is experimental. This study tests open source LLMs on their narrative generation in different genres, using ChatGPT [1] as an evaluator. Additional details are provided in section III.

This section assesses whether open-source LLMs are capable of generation comparable to GPT-4, giving a more cost-effective alternative for game designers. Because LLMs often include content moderation systems to ensure user safety, it is hypothesized that these models may demonstrate reduced performance in gruesome genres, such as horror.

2 Literature Review

This section presents past research relevant to the research question and how they relate to the current study. All cited works are listed in section VII. This section is organized into three subsections using a topical approach. Topics include: generation using LLMs (II.I), real life applications of LLMs in video games (II.II), and methods of evaluating LLMs (II.III).

2.1 Generation Using LLMS

This section is a general explanation of LLMs and their effectiveness in game generation. It has been included for some extra background knowledge. Two papers are reviewed in this section, but there are many more that explain this topic.

The first paper, Game Generation via Large Language Models[2], reviews game generation in video games using large language models including Claude, Gemma and GPTs. While it does not align with narrative generation, it shows the benefits of LLMs in video game environments.

According to Hu & Zhou et al., procedural content generation (PCG), when combined with machine learning, is limited to platformer games such as Super Mario Bros, which negatively affects its usage for games of different genres.

Procedural content generation (PCG) in video games refers to the algorithmic creation of game content such as levels, environments, characters, quests, or dialogue without direct human input for each individual element. By using rules, randomness, and sometimes artificial intelligence, PCG allows developers to generate vast amounts of varied and dynamic content efficiently, reducing development time and enhancing replayability. This proves the relevance of this paper, which aims to highlight these limitations in LLMs and to provide methods to improve these models. However, it remains important to keep in mind that this paper is in the context of narrative generation, while Hu, Zhou, et al. consider game rules and levels. Therefore, this paper does not tackle the issues they mention.

However, Hu, Zhou, et al. suggest a solution of their own. They experiment with a framework which gives both rules and the level at the same time, preventing hallucinations and allowing the context to remain constant. This proves to be a great solution to the problem, with positive results in the experiment shown in the paper. While the previous paper reviews procedural generation, the second paper, Towards grounded dialogue generation in video game environments[3], discusses narrative/dialogue generation, which is aligned to the goal of this paper. Akoury et al. use a dataset from the video game Disco Elysium: The Final Cut to prove their claim that video games can be an immersive setting for interactive story-telling.

Akoury et al. use a dataset from the game which includes vivid portrayals of many in-game entities, including actors, items, variables, and individual conversations. To aid the experiment, the game developers have clear descriptions for each of these entities, allowing a better overall understanding.

For the experiment, Akoury et al. use metrics to evaluate the dialogues generated by two LLMs, GPT-3 Curie and Codex. These metrics evaluate the overlaps of the generated dialogues with the dialogue nodes in the dataset, showing that Codex is generally better. Surprisingly, however, Codex tended to copy from the prompt twice as often as Curie and lacked cohesive plot completions due to a lack of historical context. Akoury et al. attribute the slight difference in performance of the two LLMs to the training data used for Codex. This highlights the importance of considering training data when testing LLMs and choosing an LLM for a game.

This study remains relevant to narrative generation as it tests dialogue generation, a wide-spread method to narrative generation in RPGs. Dialogue remains the heart of story-telling, and therefore is important to exploring the capabilities of LLMs in video games. That being said, this paper does not include dialogue generation, so it cannot be used as a resource for such.

2.2 Real Life Applications of LLMS in Video Games

This section includes some noteworthy examples of LLMs being used in video games in the past. It is important to know this as it can prove how well LLMs can function as generators, as well as highlight their shortcomings.

The first paper, The 2010 Mario AI Championship: Level Generation Track[4], discusses The Level Generation Competition, one of the first PCG competitions held in the world. As part of the IEEE CIS-sponsored 2010 Mario AI Championship, competitors entered level generators made for their playing preferences in Super Mario Bros, a popular mid-eighties video game.

To score, the competition focused on the enjoyment levels of various players present at the event. Judges were given a level to determine their skill level, which was then given to two level generators to evaluate. The level generators were expected to then produce a level for the judge to play. After playing each level, the judge would rank the two levels based on how fun it was and the generators would be ranked accordingly. A database and questionnaire were also used to avoid bias and ensure each judge evaluated each pair of generators at least once. While this method is effective, it does not account for the inherently subjective nature of fun, nor does it consider how this perception may fluctuate depending on a player's mood, circumstances, and other contextual factors.

The competition results show that player enjoyment over six factors in each level, including coins, rocks, enemies, spatial diversity of enemies, powerups, gaps, gap width, and spatial diversity of gaps varies for each generator, regardless of their score. This proves that the individual factors of the generated level do not determine its value, but rather how they correspond to the level itself. This conclusion connects to narrative generation as well, with the individual components of a story (characters, setting, plot, etc.) being less relevant to enjoyment compared to the story as a whole. This must be kept in mind when using a LLM for story generation, as it could change the whole course of the game.

The next paper, Collaborative Quest Completion with LLM-driven Non-Player Characters in Minecraft[5], focuses on an analytical approach to testing the effectiveness of large language

models as NPCs (non-playable characters). Through the use of Minecraft, a popular video game released in 2008, LLMs are tested on their effectiveness and their interactions with 28 different players.

To start the prompt, the NPCs were given a description of the setting of the minigame and a quick summary of the storyline. This was to introduce the scenario to prevent confusion for GPT-4. This is a necessary step when dealing with LLMs as they can use context for their responses and dialogues, and is highly effective at preventing thematically inappropriate content from generating. To further strengthen GPT-4's understanding, each NPC was given their character's backstory and situation. Rao et al. state that this helps them have an agency of their own' which allows for richer tones and language used in the game. It solidifies the character's emotions and motives, allowing the player to feel the intensity of the story and its devices. These factors help the NPCs fit into their background, a village in Minecraft. To enforce this, the NPCs were also taught basic skills used in the game to fit their character. While this is ideally effective, it proved to be an issue for the NPCs to avoid failures in their actions. To combat this, they were given a dialogue response to each type of failure using a function return value'. NPCs were also told their constraints to allow them to tell the player what they could and could not do. For example, Elena was not able to fight in the minigame. GPT-4 was told this as a prompt and therefore successfully told the player as such. Highlighting the character's limitations alongside their abilities therefore proved highly effective for the LLM, allowing it to avoid errors.

To prevent deviation from the plot, or the NPC being forced to answer questions it does not know, GPT-4 was given a subgoal generation' feature allowing it to bring the player back on track. After 6 question and response turns, GPT-4 would generate a subgoal for the NPC to return the conversion to its original topic. This is important as it prevents guesses from the LLM from being used as responses to a player's question that may not be included in the prompt given by the designer. As seen with the backstory, it prevents thematically incorrect generation and avoids confusing the NPC.

The next part of the paper goes over the results and analysis of the experiment. Two sections - communicating in-game errors and post-study results - are especially relevant. Specifically, these are sections 3.3 and 3.6 respectively. Section 3.3 discusses the NPCs' responses to players calling out their errors. If the LLM/NPC fails to do a task, it quickly apologizes to the player or asks for their help with the task. By doing so, the LLM is able to avoid logical fallacies and rewrite their behavior. This method is a great fail-safe in case the LLM does generate incorrect content and can quickly recover the situation.

The other section found relevant was section 3.6, which presents the players' responses to the post-study survey they were given after playing. Players were asked two questions: what they liked most, and what they liked least. For the first question, most players enjoyed how the AI made the experience more vivid and realistic, while others stated they were able to play better with the AI's help. These two points highlight the effectiveness of LLMs in this experiment and how they could further benefit others. On the other hand, when asked the second question, many players struggled with the LLM's lack of understanding of the environment, often leading to errors and issues with communication. One solution to solve this problem could be better fine-tuning of the prompt to allow a more thorough analysis of the world. This could be nothing more than a quick description, but would change the outcome significantly. This is relevant to the aims of this paper as it highlights key elements of game development that are of difficulty for LLMs, which is important for game developers to consider as they use this technology for their games.

The third paper, Language as Reality: A Co-Creative Storytelling Game Experience in 1001 Nights using Generative AI[6], highlights the use of GenAI to reimagine a popular folklore, 1001 Nights, in video game format. While playing the game, players can materialize spoken objects to aid their fight with the King, the antagonist. The King and storylines are purely LLM based, creating an innovative experience.

Sun et al. define an AI Native game as one where GenAI is not just an added feature but is fundamental to the game's existence and mechanism. This idea is fundamental to understanding LLMs, especially GenAI, in video games, as it clearly defines the importance the LLM is given upon implementation. It frames GenAI as a separate entity, the core of the game itself. They also mention that this new form of AI in games can lead to new genres and offer a whole new perspective to procedural content generation (PCG).

One issue their mechanism faced is the lack of control over the king's dialogue. If the king mentions something thematically incorrect, it can destroy the authenticity of the game's environment and create confusion in the narrative. This is one of the major drawbacks of using GenAI in gaming. This can be avoided using prompt engineering to provide the most information possible to the LLM.

The fourth and final paper, Procedural Artificial Narrative using Generative AI for Turn-Based Video Games[7], introduces and explores a new software, PANGeA, which aids in using LLMs for narrative content in turn-based role-playing games (RPGs). This discovery could be revolutionary for procedural generation. PANGeA uses multiple components to work, including a memory system, validation system, LLM interface, REST API, and a Unity game engine plug-in.

PANGeA uses these components to control the LLMs and their generated narratives. For example, the memory system stores past data and tokens, as well as quick summaries of the generated narratives in the game. It aids in character interaction and dialogue in a way never seen before. This method to prevent

thematically inappropriate generation is highly effective and extremely useful for game designers. Using this, it is no longer necessary for the designer to constantly input repeated prompts as PANGeA's memory system does it for them[7].

PANGeA is used in a game, Dark Shadows, which is a detective game with LLM-driven NPCs. This game proves the effectiveness of PANGeA's mechanics, with human-like NPCs with their own personalities. PANGeA uses its memory system to answer interrogations and raise suspicions for a character, making the game more enjoyable for the player. This software is an example of a possible future direction for the research in this paper.

2.3 Methods of Evaluating LLMS

This section uses past papers to reference different evaluation methods used on LLMs, and the results acquired. A similar method is used in this paper, with a quick experiment using ChatGPT[1] as an evaluator.

In the first paper, Constitutional AI: Harmlessness from AI Feedback[8], AI is evaluated using other AIs. To improve itself, AI gets trained to supervise themselves, without any human interference. A list of rules is provided to aid the process, which the AI assistant uses to identify incorrect and harmful outputting. The assistant then acts accordingly, explaining the problem and intended response to the AI it's supervising. This method of evaluation makes error identification more precise and effective than the human eye, allowing better AI systems overall.

The process is divided into two parts, the first being the Supervised Stage'. This stage includes the training of the assistant using harmful outputs. The model is asked to respond to a harmful statement, and then evaluate its response. The second part is the RL Stage'. This uses RLHF methods to input the AI feedback in place of human feedback. The feedback is then sent back to the AI assistant through an LM model. The AI assistant uses the revised feedback as a comparison for its responses to improve them.

This method of evaluation is very helpful for humans as it decreases time, cost, and energy used for evaluating LLMs. It proves that AI can judge other AI, and effectively prevent harmful content from generating. This supports the objectivity of the approach used in this paper, justifying the use of ChatGPT to evaluate the LLMs.

In the next paper, Beyond the Imitation Game[9], LLMs are tested using the BIG-bench' to thoroughly evaluate their capabilities. This allows game developers to understand the ever-growing list of abilities LLMs have, and helps them better understand future advancements in the field. BIG-bench' stands for Beyond the Imitation Game benchmark[9], and includes 204 different tasks for LLMs to complete. These tasks come from a wide range of topics, including math, english, biology, physics, chemistry, and more. These tasks were also completed by a panel of human judges to set the proper benchmark.

Because these tasks were outside the LLMs' capabilities, most LLMs performed poorly. However, they improved over time, with similar results over different categories of LLMs. These LLMs include the GPT series and some of Google's models. This experiment shows a case of human evaluation of LLMs, and how LLMs failed to reach this benchmark. While the results are not ideal, they prove that LLMs have a huge room of improvement, allowing for many advancements in the future. This method of evaluation effectively conveys the growth ability of LLMs and is an excellent example of human evaluation for LLMs. However, this method is not used in this paper due to computational constraints, and is instead a possible future route for this research.

The method used for evaluation in this study is ChatGPT. To assess its validity as a writing evaluator, the article Assessing ChatGPT's Writing Evaluation Skills using Benchmark Data by Baffour et al. is examined. The article compares ChatGPT's scoring of essays with human evaluation using the standard metric of quadratic weighted kappa, thereby reinforcing the reliability of its findings.

The study concludes that ChatGPT tends to be more lenient than human evaluators, functioning as an easy grader. It is important to note, however, that these findings are based on the PERSUADE dataset and the ASAP benchmarkneither of which is used in the present study. Despite this limitation, the findings remain relevant, as they underscore a potential drawback in using ChatGPT as an evaluator. Nonetheless, due to its accessibility and consistency, ChatGPT remains the most practical tool for this research.

3 Methodology

The experimental methodology employed in this study is straightforward. To gain a thorough understanding of the capabilities of LLMs in narrative generations, LLMs were prompted to generate a plot line for a specific genre. This was repeated for a total of six LLMs and eight genres. This method captures a wide spectrum of narrative approaches a game designer could use for their game, and to make this research more relevant. The LLMs used were gemma2 9B, llama3 8B, llama2 7B, llama2 uncensored 7B, wizard-vicuna 13B, and wizard-vicuna uncensored 7B. These LLMs are all based off of llama, and use similar frameworks. These LLMs are among the most widely used open-source models currently available. Uncensored models available for testing were also included to compare to the censored models. This is significant for the more graphic genres (such as horror), as they can be censored in the original LLM. The genres I included were fantasy, non-fiction, sci-fi, horror, romance, mystery, comedy, and adventure. These genres are some of the most popular genres currently used in both video games and story-telling, and must be effectively modeled for

the LLMs to use in narrative generation. Using this testing data, each LLM was asked to generate a plotline relevant to the genre included in the prompt. The exact prompt used was: **Can you create an engaging plot line for a [genre] story? Please include a detailed beginning, middle, and end, as well as key characters and major plot twists. The story should have a clear conflict and resolution.**[1] Once the prompt was issued, the LLMs' responses were submitted to ChatGPT (GPT-3.5 model) for evaluation. The exact prompt used for ChatGPT was: **Can you rate this story as a [genre] on a scale of 1-100?** The LLMs were graded on 6 categories: Concept & Originality, Plot Structure & Pacing, Characterization & Development, Setting & Atmosphere, Language & Style, and Emotional/Thematic Depth. The points obtained for each of these categories were compiled and presented as an overall score. ChatGPT was used for scoring because according to past research, it can mimic humans almost as accurately as GPT-4 (see section II.IV for exact research review). ChatGPT was asked to score on a scale of 1-100, and this was repeated five times for each response. After repeating, the average score was recorded in a table (see section IV). This repetition process helps when using LLMs when scoring because it helps reduce variance. Future studies may consider using GPT-4 when recreating.

4 Results

As outlined in section III, narratives generated by different LLMs were evaluated by ChatGPT. Each LLM was tasked with generating eight narratives, one for each genre. The evaluation scores are presented below.

Table 1 Table including all scores given by ChatGPT. It includes the average scores for each genre and LLM. (see fig 2 and fig 3 for graph)

		Large Language Models						
		Gemma2	Llama3	Llama2	Llama2 uncensored	Wizard-vicuna	Wizard-vicuna uncensored	**Averages**
Genre	Fantasy	85	85	**90**	78	74	75	**81**
	Non-fiction	52	**88**	40	60	72	45	60
	Sci-fi	**90**	85	89	70	75	67	79
	Horror	80	**83**	81	64	68	72	75
	Romance	**87**	85	81	75	73	72	79
	Mystery	**85**	**85**	80	66	74	64	76
	Comedy	**83**	74	75	43	70	50	66
	Adventure	85	**91**	88	77	80	58	80
	Averages	81	**85**	78	67	73	63	

This data was converted into charts for comparison. The charts are presented below

ChatGPT's scoring was relatively consistent. As mentioned in the methodology section (section III), ChatGPT gave each narrative five scores, from which an average was calculated.

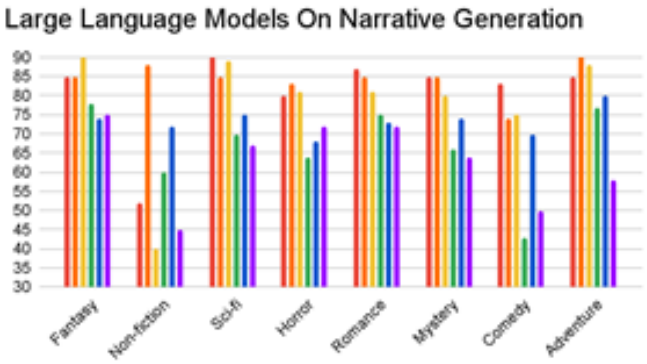

Fig 1.1 This graph is a representation of table 1 (see above). It includes each LLM's score on each genre using color coding to create vivid comparisons.

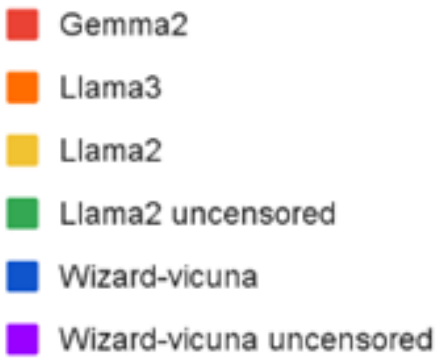

Fig 1.2 Key for figure 1.1 Each color is a different LLM.

For example, Wizard-Vicuna uncensored's horror narrative got the following scores:

Table 2 ChatGPT's scores for Wizard-Vicuna uncensored's horror narrative

75	70	75	70	70

Similarly, for Gemma2's adventure narrative, ChatGPT gave the following scores:

Table 3 ChatGPT's scores for Gemma2's adventure narrative

85	85	85	85	85

As shown, ChatGPT's scoring has low deviation, indicating the stability in the scoring data.

The lowest performing large language model was wizard-vicuna uncensored, with an average score of 63. It was followed by llama2 uncensored, with an average score of 67. Contrary to the original hypothesis, both uncensored LLMs performed the most poorly out of all six large language models. This is

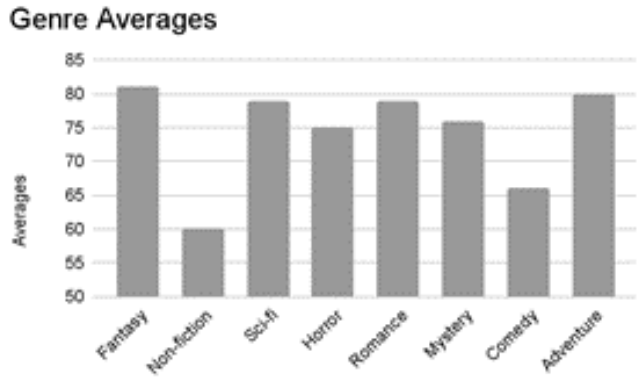

Fig 2. Average performance on each genre, represented using a bar graph. This allows easy comparison. As shown, non fiction had the lowest average, while fantasy had the highest

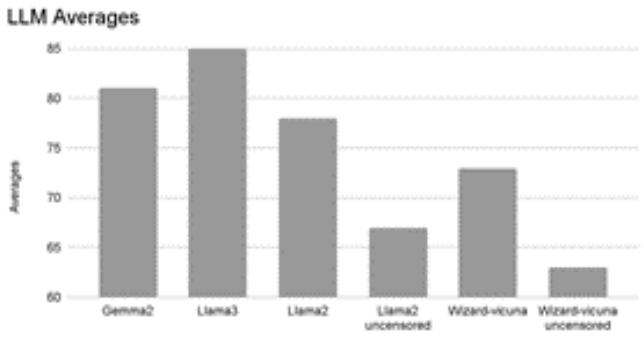

Fig 3. Average performance of each LLM over the eight genres, represented using a bar graph. As shown, wizard-vicuna uncensored had the lowest average, while llama3 had the highest

because their answers were notably brief and lacked coherent structure, unlike the others, which had strict structure and made their stories easier to understand. Additionally, both LLMs exhibited repetitive openings across genres, with minimal variation between narratives. This is likely a result of modifications made during the uncensorship process, which may have impacted narrative coherence or output diversity.

Moreover, the uncensored LLMs demonstrated difficulty with adhering to genre conventions. For example, llama2 uncensored received a notably low score for comedy due to a tone that was overly earnest and insufficiently humourous. The horror narrative lacked tension and suspense, instead emphasizing emotional themes that rendered the story more tragic than frightening. For example, Llama2's horror narrative contained this except: "Mara entered the old cabin. The lights flickered. She heard something... but it was just the wind. She went home and had dinner." This avoids graphic content, leaving the horror narrative tensionless. The LLM with the highest average score was llama3, with an average of 85. This is because it demonstrated adherence to genre expectations and employed coherent narrative structure, character development, and thematic depth. For example, an excerpt from Llama3's highly scoring fantasy narrative is presented below.

"The dragon wasn't guarding goldit was guarding a forgotten language. To speak it meant rewriting fate. Elara, a farmgirl fluent in dreams, was the only one who could." This narrative scored highly for originality, structure, and emotional resonance. It shows the model's fluency in common fantasy tropes and creativity within genre conventions.

Overall, all LLMs produced thematically comparable outputs across genres, and evaluation scores were more strongly influenced by narrative style and structure than by thematic originality. This means that the more advanced LLMs, such as llama3 and gemma2 achieved the highest scores across genres, while low-performing LLMs (such as the uncensored ones) consistently received lower scores.

All the LLMs received the lowest score in the nonfiction genre. Looking at ChatGPT's feedback, this was attributed to fictional elements being included in the responses, which diminished the factual integrity of the responses. For example, an excerpt from Llama2-uncensored's non-fiction response is shown below: "In the year 1812, dragons roamed the battlefields alongside Napoleon's armies" The inclusion of dragons makes the story overtly fantastical, removing the non-fictional integrity of the response.

The genre with the highest score, meanwhile, was the fantasy genre, likely due to the prevalence of fantasy content in LLM training data. To test this hypothesis, publicly available training data sources were analyzed. This includes a random selection of public domain books, user-generated fiction platforms, and other available literature. From the sample size, analysis indicated that 44% of training data available for LLMs was fantasy. No non-fiction works were identified within the analyzed sample, but 24% of the books were historical fiction. This means they contained a mix of fictional and non-fictional elements, similar to the narratives produced by the LLMs. These findings suggest that the genre distribution within training data likely contributes to the observed disparities in genre-specific narrative quality, particularly the underperformance in nonfiction and the relative strength in fantasy.

This is also shown when conducting statistical analysis on the results. The data is presented below.

Consistent with expectations, Llama3 received the highest mean and median score among all models. However, it did not have the lowest standard deviation, having more than Wizard-vicuna. This means that there is a greater variability in Llama3's responses compared to Wizard-vicuna. This finding is significant as higher standard deviation suggests increased inconsistency, potentially affecting its generative qualities.

For the genres, the non-fiction genre had the lowest average and the highest standard deviation. This indicates the LLMs'

Table 4 Statistical Analysis on LLM results across genres

LLM	Mean	Median	Standard Deviation
Gemma2	81	85	11.2
Llama3	85	85	4.6
Llama2	78	81	15.2
Llama2-uncensored	67	68	10.8
Wizard-vicuna	73	74	3.3
Wizard-vicuna uncensored	63	66	10.3

Table 5 Statistical Analysis on Genre results across LLMs

Genre	Mean	Median	SD
Fantasy	81	82	5.9
Non-fiction	60	56	16.4
Sci-fi	79	80	9.1
Horror	75	76	7.1
Romance	79	78	5.8
Mystery	76	68	10.8
Comedy	66	72	14.3
Adventure	80	83	10.8

difficulty generating consistent outputs within this genre, as it had the highest uncertainty (16.4).

These results contradict the initial hypothesis present in section I, and suggest that, despite censorship, LLMs can perform effectively on gruesome genres such as horror. Instead, they demonstrate difficulty with nonfiction, producing content inconsistent with factual expectations. This represents a case of genre-inconsistent generation. This issue could potentially be avoided by refining the input prompt to specify genre constraints more explicitly. However, such prompt refinement was beyond the scope of the current methodology. Additionally, the uncensored LLMs performed contrary to expectations. They demonstrated a marked decline in narrative coherence and qualities, contrasting sharply with the initial expectations.

5 Limitations and Future Work

The primary limitation of this study was using ChatGPT. While it is a practical alternative for researchers with limited computational resources, when experimenting, GPT-4 is considered a more reliable evaluator due to its closer alignment with human judgment. Future replications of this study may benefit from this change. Because of this issue, most of the results are in the 75-90 score range as ChatGPT tends to assign scores clustered around 85. Moreover, ChatGPT is a censored model, and therefore may exhibit preferential bias toward similarly moderated models. As discussed in section II.III, prior research has indicated that ChatGPT may not serve as the most effective standalone evaluator for assessing creative writing. To enhance evaluative rigor, future studies are encouraged to pair automated assessment tools with human evaluators, particularly in studies involving subjective narrative elements.

Another limitation of this study was the inability to use human responses. This study lacked access to a sufficiently large sample of human evaluators to conduct a statistically robust assessment, and therefore necessitated reliance on automated evaluation through ChatGPT. For future research, it is suggested to use human scoring to maximize usefulness of results.

A final consideration is the reliability and generalizability of this study. While it covers narrative generation thoroughly, its findings should not be considered universally applicable. The genres covered in this experiment represent only a subset of the broader range of storytelling genres, and researchers are encouraged to consult a variety of sources to gain a more comprehensive understanding. Moreover, the availability of open source large language models is continually evolving, making comprehensive coverage beyond the scope of this study. It is also important to consider the parameter number of the LLMs used (see section IV). Due to resource constraints, the study utilized LLMs with relatively low parameter counts. However, higher parameter versions of these models may demonstrate improved performance on more capable hardware.

Given that this research is purely quantitative, further research could include qualitative analysis as well. Moreover, more variation within the evaluators could also be implemented (such as human evaluators). Another route for LLMs in video games could be to focus on other aspects in gaming, including gameplay and character design. To further refine research in this field, models can be developed and trained to identify issues within AI-generated narratives.

This research can be used by game developers to enhance the experience of their games. Using LLMs for narrative generation allows for more immersive story experiences and can increase player engagement. Moreover, it enables expansive narrative possibilities, including the potential for branching or multiple storylines within a single game environment. Also, the results of this paper can be generalized into the broader field of LLMs in video games as seen in other works in this field. For example, a LLM, GENEVA, built for narrative generation in video games noticed similar drawbacks to ChatGPT's analysis [10].

6 Conclusion

This paper investigates the capability of large language models in the context of narrative generation. It aims to answer the question, How can LLMs benefit narrative generation in video

games, and what are the strategies for preventing generation of harmful or thematically inappropriate content while doing so? The study experiments with widely available and popular open-source large language models to evaluate their performance with different genres in narrative generation. While a lot of literature reviews and papers have considered procedural content generation, there are few that test narrative generation specifically. This paper seeks to address that gap, offering insights that may inform the use of language models in dynamic and genre-sensitive storytelling for interactive media.

This paper demonstrates that AI is capable of being an essential tool for narrative generation of video games, with a few exceptions (such as in genres with non-fictional elements). It is important to note that human-authored narratives currently surpass LLM outputs in coherence, emotional nuance, and genre control. While LLMs have the ability to write good stories, they tend to confuse genres and produce inconsistent narratives, a limitation less observed in human-authored works. Nonetheless, the experiment suggests that LLMs hold significant potential to assist game designers in enhancing narrative elements.

References

1 OpenAI, *ChatGPT [Large Language Model]*, https://chatgpt.com/, 2024.

2 C. Hu, Y. Zhao and J. Liu, 2024 IEEE Conference on Games (CoG), Milan, Italy, 2024, pp. 1–4.

3 N. Akoury, R. Salz, M. Iyyer and University of Massachusetts Amherst, *Towards grounded dialogue generation in video game environments*, https://people.cs.umass.edu/~nsa/papers/discoelysium_aaai_2023.pdf, 2023.

4 N. Shaker, J. Togelius, G. Yannakakis and R. Baumgarten, *ResearchGate*, 2011.

5 S. Rao, W. Xu, M. Xu, J. J. G. Leandro, K. Lobb, G. DesGarennes, C. Brockett and B. Dolan, *Collaborative Quest Completion With LLM-driven Non-Player Characters in Minecraft*, Microsoft Research, 2024, https://www.microsoft.com/en-us/research/publication/collaborative-quest-completion-with-llm-driven-non-player-characters-in-minecraft/, Published September 17, 2024.

6 Y. Sun, Z. Li, K. Fang, C. H. Lee and A. Asadipour, *arXiv*, 2023.

7 S. Buongiorno, L. Klinkert, Z. Zhuang, T. Chawla and C. Clark, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2024, pp. 156–166.

8 Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukoit, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. Dassarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. J. Henighan, T. Hume, S. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. B. Brown and J. Kaplan, *arXiv*, 2022, **abs/2212.08073**, year.

9 A. Srivastava, A. Rastogi, A. Rao, A. a. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv and Z. Wu, *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*, OpenReview, 2023, https://openreview.net/forum?id=uyTL5Bvosj&nesting=2&sort=date-desc.

10 B. Potts, *GENEVA uses large language models for interactive game narrative design*, Microsoft Research, 2024, https://www.microsoft.com/en-us/research/blog/geneva-uses-large-language-models-for-interactive-game-narrative-design/, Published August 5, 2024.

11 Google, *Gemma 2 [Large Language Model]*, https://ai.google.dev/gemma?gad_source=1&gclid=Cj0KCQjwwuG1BhCnARIsAFWBUC06b5AX3HwaNYYLcJGCxKjaDfj-cNeFlT6TUFDlYY9Zn6p825Ves6YaAp9vEALw_wcB, 2024.

12 Meta, *Llama 3 (April version) [Large Language Model]*, https://llama.meta.com/, 2024.

13 Meta Platforms, *Llama 2 [Large Language Model]*, https://llama.meta.com/docs/model-cards-and-prompt-formats/other-models#meta-llama-2, 2024.

14 G. Sung and J. Hope, *Llama 2 uncensored [Large Language Model]*, https://huggingface.co/georgesung/llama2_7b_chat_uncensored, 2024.

15 Junelee and MelodysDreamj, *Wizard-vicuna [Large Language Model]*, https://huggingface.co/junelee/wizard-vicuna-13b, 2024.

16 E. Hartford, *Wizard-vicuna-uncensored [Large Language Model]*, https://huggingface.co/cognitivecomputations/Wizard-Vicuna-7B-Uncensored, 2024.

17 Z. Luo, Q. Xie and S. Ananiadou, *arXiv*, 2023.