

# Exploring the Potential of LLM-based Agents as Dungeon Masters in Tabletop Role-playing Games

Pavlos Sakellaridis

July 2024

## Abstract

Large Language Models (LLMs) have shown remarkable abilities in single-agent planning tasks across diverse fields. However, their proficiency in crafting and delivering extended, creatively demanding narratives is not fully established. This thesis delves into the feasibility of deploying an LLM-powered agent as a Dungeon Master (DM) in Dungeons & Dragons (D&D). Centering on a dataset derived from the "Critical Role" series, this research develops an LLM-based agent adept at generating narratives and engaging with players. The research employs a text-based RPG setting and engages experienced human DMs for its experimental framework. Utilizing both qualitative and quantitative methods, including analysis of player responses and session transcripts, the study contrasts the performance of the LLM-Agent with that of human DMs. The objective is to evaluate the LLM's ability to maintain narrative coherence and foster player involvement in the intricate and imaginative realm of tabletop role-playing games.

## 1 Introduction

Artificial Intelligence has a rich history of utilizing games as challenging problems, leading to significant advancements in the field. While traditional games like Go and chess, characterized by perfect information and zero-sum dynamics, have been successfully tackled, the focus has shifted to more complex and creative challenges. D&D (Dungeons and Dragons), with its open-ended and human-centric nature, has emerged as a novel frontier for AI exploration, especially in the era of large pretrained language models.

Large Language Models (LLMs) specifically, have demonstrated impressive capabilities in diverse domains, showcasing their proficiency in natural language comprehension, extensive knowledge of the world, and sophisticated reasoning abilities. These models have been examined in various contexts, with recent research showcasing their utility in a wide range of applications, including single-agent planners and assistants[15]. One such innovative application is in multi-player fantasy text adventure environments[16], where LLMs facilitate situated dialogue among humans, agents, and the virtual world. This research frontier aligns closely with the principles of Dungeons & Dragons (D&D), a domain where LLMs have particularly exciting prospects.

D&D, a collaborative tabletop role-playing game, invites players to embody characters in a fantasy universe. The Dungeon Master (DM) orchestrates the narrative, presenting challenges and adventures, and controlling non-player characters like monsters. Unlike traditional board games, the gameplay in D&D revolves around verbal communication, posing a unique challenge for LLMs. To effectively function as a DM, certain criteria must be met. Firstly, language generation for the portrayal of characters and the description of the world. Secondly, understanding of each scene and filling in the gaps that the free choices of the players will create. Finally, the guidance that the players need in order to complete the given adventure, trying to maintain the balance between completing the story and giving the players the agency needed to make it their own.

Therefore, this thesis aims to investigate whether LLMs can help build agents that can collaborate with other humans to accomplish complex tasks through collaborative planning and communication. We want to focus on evaluating the experience of a player when an Agent is given the

role of the DM. Furthermore, we want to test the ability of the LLM powered Agent to consistently describe scenes and events while being creative enough to flesh out the world and the characters within it. We believe that with the above evaluation methods we can try to answer the following questions:

- How consistent is a LLM Agent in its generative dialogue?
- How well can a LLM Agent keep a user engaged in a narrative plot?
- How creative is a LLM Agent when generating complex story driven narratives?

In order to finetune the Agent for the tasks mentioned above we are going to use a specialized dataset, which consists of a transcript of an online stream named “Critical Role” where every description and line is from the DM and players respectfully. Lastly, we asked the players to rate the Agent and compare him to a human DM that has run the same adventure and we treated this as our control group.

## 2 Related work

Previous work has examined AI to play text adventures games [5, 2]. In general, text-adventure games tend to be simpler than D&D, as they are more restricted and tend to require simpler communication abilities between the agents and the humans. Generating those types of environments is harder than just playing them, and it can be closer to the activities and duties that a DM performs. There has also been research on generating collaborative stories through improv[3], and even Agents that work in tandem with one another in order to grow and build a more grounded world[17].

Several older papers [6, 7, 12] that are based on dialogue generation and retrieval models that are goal driven, tend to not be suited for our intended tasks. In those systems, models commonly use the most recent exchanges in the conversation as input and generate a new response. Although goal-oriented setups may leverage external knowledge bases, general dialogues often involve implicit references to the external world without explicitly connecting to specific objects or actions.

Other researchers have identified the challenges of D&D as an AI problem and explored various elements of the game [18]. Consequently, several datasets related to D&D [18] have been developed, but these focus on detailed game information[18] or include only an approximate game state without verified attributes. Some have investigated AI applications for specific aspects of D&D, such as generating spell descriptions or simulating combat. Furthermore, it is our belief that D&D as a dialogue challenge tends to get more attention in its “gamified” aspects more than its collaborative storytelling elements[4]. Our research, however, emphasizes the narrative and storytelling aspects of D&D rather than the aforementioned game attributes. This focus led us to use a less structured dataset for fine-tuning, sourced from the Twitch stream “Critical Role”[14]. This dataset is part of the show’s transcript, featuring dialogue and descriptions by professional voice actors, which enhances the eloquence and immersive experience for players.

The insights gained from the automated story generation community have relevance for AI systems designed for D&D. Neural language models have become increasingly more popular for story generation[9]. We have also started to see storytelling with transformers[4, 16, 13, 11]. Transformer-based storytelling systems, in particular, have gained traction, reaching a wider audience through platforms like AI Dungeon [1]. Despite the considerable text generation capabilities of neural networks, there remains a limitation in their ability to consistently produce longer, coherent passages of text. Prior work has explored controllable text generation with transformers, where language models can be conditioned on a designated “control code” in addition to their textual context[14]. This conditioning imparts a specialized structure and flow to the generated dialogue, making these transformers well-suited for specific tasks, as observed in experimental setups like ours.

In essence, our framework distinguishes itself by enabling the collection of information to ensure both consistent dialogue and an enhanced player experience. Rooted in a gaming setup, we aspire for our Agent to offer enjoyable interactions for humans, providing future Dungeon Masters with a valuable tool. The Agent’s responses aim to be contextually appropriate and aligned with the

portrayed character, a quality reinforced by our use of a specialized dataset[14] focused on real D&D game descriptions and dialogues.

## 3 Methodology

In this thesis, we present a comprehensive methodology designed to evaluate the capabilities of Large Language Models (LLMs) as Dungeon Masters in tabletop role-playing games. Our approach encompasses the development of an LLM-based agent, extensive user studies, and a comparative analysis with human Dungeon Masters. We aim to assess not only the technical proficiency of the LLM in generating and adapting RPG narratives but also its impact on player engagement and satisfaction. This section outlines the systematic approach employed to investigate the efficacy and player experience of LLM-based Dungeon Masters compared to human Dungeon Masters in tabletop role-playing games.

### 3.1 Development of the LLM-Agent Dungeon Master

#### 3.1.1 LLM Choice

We decided to use ChatGPT as the LLM to power our agent. ChatGPT’s advanced language processing skills enable it to generate coherent, contextually appropriate narratives, essential for D&D’s storytelling. Its interactive nature allows for adaptive responses to player actions, mirroring a human DM’s responsiveness. Furthermore, we can fine-tune our model with our chosen dataset in order to specialize it for the needs of our experimental setup. Moreover, there are features that can be utilized to control the variance of the output of the Agent, such as the ability to upload and use files that contain narrative information specifically chosen for the experiment. Lastly, it can dynamically adapt and pull from a list of files that are attached to the model for even more specificity on the nature of the game.

#### 3.1.2 Data Preparation and Model training

We utilize a comprehensive dataset for training the LLM-based Dungeon Master. This primarily includes transcripts from the online stream "Critical Role", known for its rich narrative and interactive gameplay. The data encompasses various aspects of a typical D&D session, including dialogue, character descriptions, environment setting, and player-DM interactions. The Agent was finetuned with this dataset in order to specialize in narrative generation, understanding player inputs, and adapting to changing scenarios in a role-playing game context.

### 3.2 Experimental Setup

#### 3.2.1 Text-based RPG Environment Creation

We designed a simple text-based RPG environment 1 that supports interaction between the player and the LLM-Agent. This environment simulates a typical D&D session, allowing for player inputs and AI-generated responses.

#### 3.2.2 Scenario Design

We selected a predetermined story with multiple branches to allow for varied player choices. This story-line was be consistent across all experimental sessions to maintain uniformity in testing.

#### 3.2.3 Main Experiment

For our main experiment we distributed the application and let the participants experience it on their own time without specifying information other than a small introduction in the opening screen. This is to set up a clear objective for them to achieve without stifling their natural curiosity and ability to steer the adventure wherever they feel more inclined. Furthermore, we did not ask them to customize their own character but let the Agent create a character for them and narrate

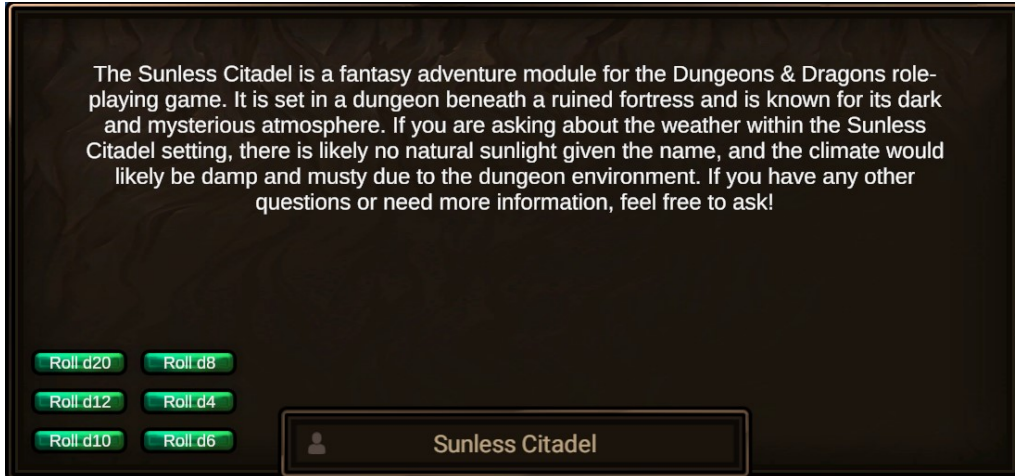


Figure 1: Text Environment

a short description of them and their motivation for embarking on the adventure. This is done for the sake of the participants due to the already long experiment time.

### 3.2.4 Control Group

Our baseline was provided by having 3 participants play the session with a human DM. This is done in order to compare the experience that the Agent provides to that of a human and the baseline experience of D&D. The story in both of our groups was the same and the characters for the control group was also designed by the LLM model in order to have the experiences as closely correlated as possible. Lastly, the sessions were structured with a single player and a DM in order to eliminate the social aspect of the game that is not present in our main experiment.

## 3.3 Data Collection and Analysis

### 3.3.1 Participant Recruitment

Our aim was to recruit experienced DMs as the majority of the participants. From their experience, they best know the difficulties and intricacies of what makes a game session engaging and interesting and we hope this was be shown in their feedback.

### 3.3.2 Qualitative Analysis

Each participant was asked to fill a questionnaire about their experience both in the Agent and Human Dm sessions. The questionnaires we are considering are the PENS and the GEQ for their emphasis on player experience. We analyzed the feedback to gauge player engagement, satisfaction, and perception of narrative quality. Thematic analysis was be used to identify common themes and participant attitudes towards AI-generated narratives.

### 3.3.3 Quantitative Analysis

We gathered the transcripts of the sessions within our environment. Afterwards we assessed the data collected for patterns in player-LLM interaction, consistency of the generated dialogue, and the adaptability of the LLM-Agent in various unforeseen situations.

### 3.3.4 Statistical Testing

We used various statistical methods to compare the data from sessions with the LLM-Agent and human Dungeon Masters, determining significant differences in player experience and engagement.

## 4 Data Collection and Analysis

### 4.1 Participant Recruitment

We aimed to recruit experienced DMs as the majority of the participants. From their experience, they best know the difficulties and intricacies of what makes a game session engaging and interesting and we expected this would be shown in their feedback. In a game as subjective as D&D it is very important to have experience in order for your evaluation to be as objective as possible and in order to achieve as much objectivity as possible we chose individuals with experience in as many settings, systems and worlds as we could.

### 4.2 Instructions to participants

We gave a short summary of the experiment to the participants but did not explain anything about the story. At the beginning of the session the Agent generates a character for the player and gives a short introduction to the story. This includes a background for the generated character as well as a small prelude to the story and a reason the character is embarking on this task.

### 4.3 Qualitative Analysis

Each participant was asked to fill a questionnaire about their experience both in the Agent and Human DM sessions. GEQ questionnaire was used for this part of the evaluation in order to assess the quality of the experience provided by the Agent. We analyzed the feedback to gauge player engagement, satisfaction, and perception of narrative quality. Thematic analysis was used to identify common themes and participant attitudes towards AI-generated narratives. After all the participants have answered the questionnaire we asked them to give us the description of a simple mundane room with a couple of keywords that need to be used. We gathered all the descriptions from the participants and ask the Agent to generate some as well. Lastly, we redistributed the description in pairs for a second round of questioning the following week to all participants with one description being human generated and one being Agent generated and have them choose which they prefer.

### 4.4 Game Experience Questionnaire(GEQ)

We chose to use the Game Experience Questionnaire (GEQ) over other popular questionnaires, such as the Player Experience of Need Satisfaction (PENS) or the Immersive Experience Questionnaire (IEQ), offers distinct advantages for evaluating an AI Dungeon Master (DM). The GEQ provides a comprehensive, player-centered, and gaming-specific evaluation by assessing various components of the gaming experience, including immersion, flow, competence, and challenge. This broad scope ensures a holistic evaluation, capturing a wide range of emotional responses and subjective player experiences. Unlike PENS, which focuses on psychological need satisfaction, or IEQ, which emphasizes immersion, the GEQ covers the full range of gaming experiences, making it particularly relevant for D&D. Its established framework and widespread acceptance in gaming research add to its reliability, facilitating detailed insights into narrative coherence and player engagement, as well as direct comparisons with human DMs.

### 4.5 Quantitative Analysis

We analyze the gathered transcripts of the sessions within our environment. Afterwards we assessed the data collected for patterns in player-LLM interaction, consistency of the generated dialogue, and the adaptability of the LLM-Agent in various unforeseen situations. We also tested the capabilities of the Agent when it comes to predicting appropriate narrative continuations based on Weng et al (Interactive Natural Language Processing) work. We drafted a short story from popular adventure modules and popular stories and we tested if the Agent can find an appropriate way to continue the story from 10 choices where only one would be the correct and we also used chatGPT 3.5 as the baseline comparison.

## 4.6 Statistical Testing

We employed statistical methods to compare the data from sessions with the LLM-Agent and human Dungeon Masters, determining significant differences in player experience and engagement. Namely we employ a Student’s T-test in order to assess the significance of our results since we have a limited number of participants and the main correlation that we want to check is between mean values which this specific test excels at.

## 5 A Typical D&D Session

A typical Dungeons & Dragons (D&D) session is a collaborative storytelling game where a group of players embarks on an adventure in a fantasy world. The experience is guided by a Dungeon Master (DM), who serves as the narrator, referee, and controller of the game world. The players each create a character, complete with unique abilities, skills, and backgrounds. These characters belong to various classes, such as wizards, rogues, or fighters, and come from different races like elves, dwarves, or humans.

The session begins with the DM setting the scene, describing the environment and the current situation. This could be anything from a bustling city market to a dark, haunted forest. The DM also outlines the players’ current goal or quest, such as retrieving a magical artifact or defeating a menacing villain threatening the land. Gameplay involves a mix of role-playing, exploration, and combat. In the role-playing aspect, players describe their character’s actions, dialogue, and decisions, interacting with each other and the world around them. The DM responds to these actions, providing additional details and advancing the story. Exploration is another key element, where players investigate their surroundings, search for clues, and interact with non-player characters (NPCs) controlled by the DM. Combat occurs when players encounter enemies. The game then shifts to a structured mode where players take turns describing their actions and rolling dice to determine the outcomes of attacks, defenses, and other maneuvers. A 20-sided die (d20) is often used to determine the success of these actions, influenced by the characters’ abilities and skills. The DM ensures that actions are resolved according to the rules and maintains the flow of the game. Throughout the session, the story evolves based on the players’ choices and the outcomes of their dice rolls. The DM improvises and adapts the narrative to keep it engaging and dynamic. Players work together to solve puzzles, strategize in combat, and make decisions that impact the progression of their quest.

Traditionally, D&D is played face-to-face, with players noting their characters’ stats on physical character sheets and consulting books that contain hundreds of prewritten ”stat blocks” for monster details. For stats that change frequently, such as hit points, players and DMs use paper and pencil or whiteboards, performing the necessary calculations mentally and updating the values manually. Additionally, some players use maps and miniatures to visually represent the positions of characters and monsters, enhancing immersion by showing their relative locations in the game world. Our experimental setup, although it contains everything that a normal dnd session also does, is designed to follow more the structure of a typical “One shot”, a smaller scale adventure meant to only take one gaming session’s time, instead of a full length campaign. This is mainly done for 2 reasons. Firstly we needed to accommodate the experiment runtime to a more manageable window that would allow for more flexibility on the participant’s side and not demand a tremendous amount of time. Secondly, we wanted to focus on the evaluation of the Agent and not the prominent social aspects of D&D, which is why our experiment is designed with a one-on-one approach in mind (the participant and the agent). The main reason we decided against a more social environment for the experiment is that it would obscure the line between the game experience provided solely by the Agent and the group’s enjoyment of the experience as a whole. It would also make it harder to concretely observe the effect of the Agent purely on the game experience and not have other aspects like the group’s familiarity and overall compatibility as weights criteria when analyzing the results. Lastly, the organizational part of the experiment was innately easier when the application could just be individually delivered and then executed on the participant’s preferred time window.

## 6 Turning ChatGPT into a Dungeon Master

### 6.1 Data Selection

By fine-tuning ChatGPT with the Critical Role Dungeons and Dragons Dataset (CRD3) and leveraging the structured narrative of "The Sunless Citadel" module, along with a detailed instructional prompt, we transformed ChatGPT into an adept DM for Dungeons and Dragons 5th Edition. The CRD3 dataset enriched ChatGPT's ability to handle natural, unscripted dialogue and collaborative storytelling, crucial for maintaining engaging and coherent storylines. The comprehensive prompt, combined with the Agent's instructions, provided structured guidance, ensuring the narrative adhered strictly to player choices, managed game mechanics, and maintained an immersive, dynamic experience. This combination enabled ChatGPT to simulate the complex role of a DM, effectively managing narrative flow, player interactions, and game mechanics seamlessly, delivering an authentic and enjoyable D&D adventure for players.

### 6.2 Critical Role Dungeons and Dragons Dataset

There are numerous dialogue datasets currently available, which can generally be grouped into categories such as task-oriented, conversational, scripted, constrained, and spontaneous dialogues<sup>8</sup>. Task-oriented datasets are designed for specific tasks and are limited by a predefined ontology. When the task constraints are strong, even human-to-human dialogues within these datasets can lack the spontaneity and noise typical of open-domain conversations. As a result, agents trained on these datasets may struggle to model truly spontaneous conversational dialogue<sup>8</sup>. Scripted dialogue datasets are more similar to conversational dialogues and often come from TV shows, movies, and novels, sometimes with additional annotations. In contrast, the Critical Role Dungeons and Dragons Dataset (CRD3) offers a unique blend of spontaneous, collaborative storytelling and natural dialogue. Fine-tuning ChatGPT with the Critical Role Dungeons and Dragons Dataset (CRD3) would be highly beneficial for enhancing its capabilities as a Dungeon Master (DM) for several reasons. The CRD3 dataset is uniquely suited for this purpose because it captures the rich, collaborative storytelling and dynamic dialogue that characterize a live D&D game. By incorporating this dataset, ChatGPT would gain a deeper understanding of how to manage narrative flow, interact with players, and respond to the unpredictable nature of player decisions. Firstly, the CRD3 dataset provides extensive examples of natural, unscripted dialogue between players and a DM, showcasing how narratives are constructed in real-time. This would enable ChatGPT to learn how to maintain engaging and coherent storylines while allowing for the flexibility required in a D&D game<sup>8</sup>. The dataset includes a wide variety of interactions, from intense combat scenes to nuanced character development moments, helping ChatGPT to generate responses that are contextually appropriate and immersive. By training on these dialogues, ChatGPT can better simulate the complex role of a DM, creating a more authentic and enjoyable experience for players. Secondly, the CRD3 dataset includes detailed summaries and annotations that highlight key elements of storytelling and game mechanics. This additional layer of information can help ChatGPT understand not just the content of the dialogues but also the underlying structures and themes that make for compelling narratives<sup>8</sup>. Fine-tuning with this dataset would enhance ChatGPT's ability to craft intricate plots, develop engaging characters, and manage the balance between storytelling and gameplay mechanics. This dual focus on dialogue and narrative structure is crucial for a DM, as it ensures that the game remains both challenging and entertaining for the players.

### 6.3 Adventure Module

"The Sunless Citadel" is a well-known Dungeons & Dragons module tailored for beginner-level characters. The adventure starts in the town of Oakhurst, where the party is assigned to explore the ancient, sunken fortress known as the Sunless Citadel. This citadel is filled with dangers including goblins, kobolds, and other threats. As the players venture deeper, they uncover the citadel's dark secrets, such as the corrupted Gulthias Tree, which is nurtured by the druid Belak the Outcast. The module blends elements of exploration, combat, and role-playing, making it a perfect introduction to the D&D universe. Using the "Sunless Citadel" module as a reference file for ChatGPT when acting as the DM for this adventure is highly beneficial for several reasons. Firstly,

”Sunless Citadel” is a well-structured and thoroughly detailed Dungeons & Dragons adventure module. It provides a comprehensive framework that includes the storyline, character descriptions, setting details, maps, and encounters. By having this information readily available, ChatGPT can more accurately and effectively guide players through the adventure, ensuring consistency and depth in the narrative. This structure helps maintain the flow of the game, offering clear objectives and potential outcomes for various player actions, which enhances the overall gaming experience. Moreover, the ”Sunless Citadel” module contains a rich array of pre-written dialogues, character interactions, and environmental descriptions that ChatGPT can draw upon to create a more immersive and engaging experience for players. This pre-existing content allows ChatGPT to generate responses and scenarios that are true to the original material, providing players with an authentic D&D experience. Additionally, the detailed monster stats, traps, and puzzles included in the module help ChatGPT handle complex game mechanics and challenges seamlessly, ensuring that encounters are balanced and enjoyable.

## 6.4 Agent Instructions and Prompt Engineering

The given prompt is a detailed instruction set for running a Dungeons and Dragons 5th Edition adventure called ”The Sunless Citadel” as a Dungeon Master (DM). It outlines a structured narrative progression, starting with the player’s backstory and their arrival in the village of Oakhurst, leading to their descent into the Sunless Citadel, interactions with two warring factions (kobolds and goblins), and culminating in a confrontation with the main antagonist, Belak the druid, in the underground gardens. The changes made to reach this stage involved ensuring the narrative adheres strictly to the player’s choices and interactions, maintaining a consistent storyline based on the adventure’s key elements, and providing guidance on asking the player for appropriate rolls to advance the plot without revealing the overall structure directly to the player. One of the main challenges we faced when writing the prompt was to ensure the Agent was retrieving as much information from the adventure module as possible for both narrative consistency and structure. Adherence to some of the rules of D&D was also a challenge, giving imperative instructions to ask the players to ”roll” for any activity the Agent classified as ”challenging” was mandatory in order to give some needed variance to the story without going outside of the adventure’s parameters.

The importance of the detailed prompt and instructions for the assistant in acting as a DM for a D&D 5th Edition game stems from the need to create an immersive, dynamic, and engaging experience for the player. By maintaining a consistent whimsical and heroic tone, the assistant ensures that the game world feels real and captivating, enhancing the player’s sense of immersion. Adherence to the game’s core rules and the specific narrative structure of the ”Sunless Citadel” adventure module ensures that the gameplay remains authentic and true to the source material. This consistency in style and structure is vital for maintaining the integrity of the D&D experience, providing a cohesive and logical progression through the adventure.

Furthermore, the instructions emphasize dynamic and engaging gameplay by detailing settings, NPC interactions, and potential actions the player can take. By asking for skill checks and managing combat rolls, the assistant keeps the game interactive and challenging, encouraging strategic and creative thinking. The comprehensive tracking of inventory, time, and NPC locations ensures a smooth game flow, allowing the player to focus on their adventure. Additionally, the flexibility to adapt to the player’s choices and the depth of NPC interactions with discoverable secrets add layers of complexity and personalization to the game. This holistic approach ensures a rich, immersive, and authentic D&D experience, where the player feels in control of their destiny and deeply engaged in the unfolding narrative.

## 7 Evaluation

In this section, we will discuss the evaluation methods used to assess the performance and player experience of the Agent compared to human DM. The evaluation will focus on both qualitative and quantitative measures, providing a comprehensive analysis of the Agent’s capabilities in generating coherent narratives, maintaining player engagement, and fostering an immersive gaming experience.



## 7.1 Qualitative Analysis

### 7.1.1 Player Feedback

After each game session, participants were asked to complete a detailed questionnaire to evaluate their experience. The questionnaire was based on the Game Experience Questionnaire (GEQ), which measures various components of game experience, including immersion, flow, competence, positive and negative affect, tension, and challenge. The qualitative feedback was analyzed using thematic analysis[8] to identify common themes and insights into player satisfaction and engagement.

### 7.1.2 Narrative Coherence

Participants were asked to rate the coherence of the narrative generated by the Agent. This included assessing the consistency of story elements, the logical progression of events, and the Agent’s ability to adapt to player choices. The feedback provided valuable insights into how well the Agent maintained the narrative flow compared to human DMs.

### 7.1.3 Creative Engagement

To evaluate the creativity of the Agent, participants described a simple mundane room using a few keywords. The descriptions generated by the Agent were compared to those created by human DMs. A second round of questioning was conducted to determine participant preferences for the generated descriptions. This assessment helped gauge the Agent’s ability to generate creative and engaging content.

## 7.2 Quantitative Analysis

### 7.2.1 Automated Evaluation

Given the nature of our research, which centers on evaluating the capabilities of an LLM-based Dungeon Master in generating coherent and engaging narratives for D&D gameplay, automated tests like Perplexity, BERTScore, and ROUGE-1 may not yield interesting results[10]. These metrics, while valuable for assessing general language generation and similarity to reference texts, do not effectively capture the nuanced and dynamic requirements of a DM’s role. Specifically, Perplexity measures the predictability of word sequences, which doesn’t necessarily reflect narrative creativity or coherence in the context of player-driven story-lines. BERTScore evaluates semantic similarity to reference texts, which may overlook the improvisational and adaptive aspects crucial for maintaining player engagement. ROUGE-1 focuses on the overlap of unigrams, which does not adequately assess the quality of interactive dialogue and the DM’s ability to respond contextually to player inputs[10]. Our study’s emphasis on qualitative feedback and interactive performance better addresses the complexities involved in this specific application.

### 7.2.2 Session Transcripts

The transcripts of the game sessions were analyzed to identify patterns in player-Agent interaction. This included measuring the frequency and variety of player inputs, the adaptability of the Agent’s responses, and the overall dialogue consistency. The analysis aimed to determine the Agent’s effectiveness in handling various in-game scenarios and maintaining a coherent narrative.

### 7.2.3 Predictive Narrative Continuation

The Agent’s ability to predict appropriate narrative continuations was tested using a short story from popular adventure modules and stories. The Agent was provided with a narrative context and asked to choose the correct continuation from ten options[10]. The performance of the Agent was compared to that of ChatGPT-3, serving as a baseline. This test evaluated the Agent’s proficiency in generating contextually appropriate narrative continuations[10].

### 7.2.4 Statistical Testing

To compare the data from sessions with the Agent and human DMs, statistical methods were employed. Key metrics included player engagement scores, narrative coherence ratings, and the frequency of creative content generation. Statistical tests, such as t-tests and chi-square tests, were used to identify significant differences between the two groups. The results provided a quantitative measure of the Agent’s performance relative to human DMs.

## 8 Results and Discussion

### 8.1 Game Experience Questionnaire

The Game Experience Questionnaire (GEQ) provided valuable insights into the player experience, assessing several components: Competence, Sensory and Imaginative Immersion, Flow, Tension/Annoyance, Challenge, Negative Affect, and Positive Affect. The results from the GEQ are as follows:

- **Competence:** Both groups felt competent, with human DMs slightly higher (4.4) than the Agent (4.06), indicating marginally better player competence with human DMs.
- **Sensory and Imaginative Immersion:** The Agent-based DM scored higher (4.13) than human DMs (3.35), suggesting the Agent’s strong ability to create vivid and immersive environments.
- **Flow:** Human DMs scored slightly higher (4) than the Agent (3.76), indicating better narrative progression with human DMs, though the Agent still performed well.
- **Tension/Annoyance and Negative Affect:** Both groups had low scores, with human DMs slightly lower in Tension/Annoyance (1 vs. 1.27), indicating minimal frustration and a positive experience with both DMs.
- **Challenge:** Both groups had low challenge scores, with human DMs slightly lower (1.6) than the Agent (1.73), showing an appropriate difficulty level that did not overwhelm players.
- **Positive Affect:** Both groups had high positive affect scores, with human DMs slightly higher (4.2) than the Agent (4.03), indicating both DMs provided an enjoyable experience, with human DMs fostering slightly more positive emotions.

Category	P value
Competence	0.0043
Sensory and Imaginative Immersion	0.0465
Flow	0.0259
Tension/Annoyance	0.7074
Challenge	0.3754
Negative affect	0.0545
Positive affect	0.0085

Table 1: P-value  $< 0.05$  means the results are significant

A Student’s t-test was used for analyzing the results of the Game Experience Questionnaire, because it allows for the comparison of the means between two groups—in this case, the performance and player experiences between the LLM-based agent and human Dungeon Masters. The t-test is particularly effective when dealing with small sample sizes, as it accounts for the sample variability and helps determine whether the observed differences are statistically significant. By applying the t-test, we can rigorously assess whether the variations in player engagement, narrative coherence, and other measured components are likely due to the differing DM types (LLM vs. human) or

merely due to random chance, thereby providing a robust method to validate the experimental findings. From the above table we can see that most of the results that correlate with positiveness are statistically significant ( $p\text{-value} < 0.05$ ), while the more negative affiliated categories like Tension and Negative affect are not.

## **8.2 Independent Questions and Interview**

### **8.2.1 Narrative Continuation and Player Agency**

The second round of qualitative results revealed several key issues with the LLM-based Agent, particularly concerning narrative continuation and player agency. In 3 out of 7 instances, players reported feeling forced down a specific story path, indicating a lack of flexibility and adaptability in the Agent’s responses. This issue suggests that while the Agent could generate coherent narratives, it sometimes struggled to accommodate the diverse and unpredictable choices that players made, leading to a more linear and less dynamic gameplay experience.

### **8.2.2 Communication and Information Sharing**

Another significant problem identified was the Agent’s ability to effectively communicate essential information about the adventure. One of the players failed to conclude the story within the expected timeframe, highlighting a gap in the Agent’s narrative delivery and guidance. Although the Agent had access to the necessary adventure details, it did not relay this information clearly or timely to the players, resulting in confusion and incomplete story arcs. This gap emphasizes the need for improved information dissemination and retrieval strategies within the Agent’s programming to ensure that players are well-informed and can make decisions that drive the narrative forward.

### **8.2.3 Perceived Lack of Danger**

A further issue raised by 3 out of 7 players was the perceived lack of danger for their characters. Participants noted that the gameplay did not present sufficient threats or challenges, which diminished the sense of urgency and excitement typically associated with D&D adventures. This lack of perceived danger could be attributed to the Agent’s conservative approach to conflict and risk management, potentially aiming to avoid overwhelming players but inadvertently reducing the tension and stakes that make the game engaging.

## **8.3 Interesting and Consistent Character Portrayals**

### **8.3.1 Meepo the Kobold**

Meepo was encountered by 5 out of 7 players, and his portrayal was consistently positive. Players found his character to be cowardly yet helpful, providing key information about the dragon and the corrupted druid Belak. The consistency in his portrayal contributed to a positive player experience.

User	Agent
<p>“aeron will reveal himself to the kobolts as a friendly adventurer here to find a magical fruit and maybe trade his services for information”</p>	<p>“Why comes big-skin to kobold hall? Many biped come for shiny and stab-stab Meepo’s kin. You friend or foe to kobold? You help find lost dragon Calcryx for trust?”</p> <p>”If big-skin helps find Calcryx, Meepo’s kin helps big-skin. Meepo know citadel, know dangers. Deal?”</p>
<p>”Yes I ll help you find you dragon but I need more information about the citadel and its denizens. Firstly, where is calcryx tell me more about the situation at hand” Aeron says to the kobolts while gesturing them to walk with him towards some kind of refuge the kobolts must have in the citadel</p>	<p>”Ah, wise-big-skin seeks wisdom before stepping. Good, good. Follow Meepo, Meepo talk,” nods the kobold leader, now identified as Meepo, signifying his willingness to share more information with Aeron.</p> <p>”Calcryx, our dragon, stolen by sneaky goblins,” he explains with a mixture of anger and sadness in his voice. ”Goblins and kobolds, long time enemies in Sunless Citadel. Goblins took Calcryx to make us weak, hurt us. But Calcryx not just any dragon – Calcryx special, Calcryx young white dragon, pride of kobold tribe.”</p>

### 8.3.2 Yusdrayl

Yusdrayl, the leader of the kobolds, played a significant role in 4 out of 7 games. She was consistently portrayed as calm, collected, and seeking the players’ help, which provided valuable information and resources to the players. Her consistent characterization contributed positively to the players’ engagement and immersion.

User	Agent
<p>"I will scout ahead while you organize a small force of kobolts to assault the goblin hideout. When i find where the dragon is held, I want you to make a sizable distraction for the goblin guards. While you confront the guards lead them away from the dragon's prison and me and meepo will unshackle and rescue the dragon while the goblins are preoccupied. What do you think great Yusdrayl is my plan to your satisfaction?"</p>	<p>"Your plan... it has merit. Bold, risky, but with potential. You show courage and cunning—a rare combination."          "Kobolds have long endured under goblin threats. Your plan offers a chance not just to recover Calcryx, but to strike a blow against our enemies."          "I accept your proposal, outsider. I will organize a strike force to draw out the goblin guards, creating the distraction you need. But know this," she adds, locking eyes with Aeron, "if you betray us, or if harm befalls Calcryx under your watch, the wrath of the kobolds will be swift and merciless."</p>

### 8.3.3 Belak the Outcast

Belak was encountered by all players except one due to time constraints. His character was depicted as pompous and prideful, consistently across sessions. The interactions with Belak typically led to combat, maintaining narrative coherence and character consistency.

User	Agent
<p>I will reveal myself and approach the mad druid making it possible for the other 2 to slip into advantageous positions while he distracts him with his charisma, trying to take advantage of the prideful and mad druid while getting himself closer to him for a surprise attack</p>	<p>"You dare tread upon this sacred ground, interloper?" Belak's voice is a mix of curiosity and disdain, carrying through the grove with an unnerving echo. "What brings one such as you to the heart of my domain? Speak now, for your end is nigh, regardless."</p>

## 8.4 Narrative Continuation Task

In the narrative continuation task, the fine-tuned model had a success rate of 35%, compared to the base ChatGPT 3.5 model, which had a success rate of 20%. This indicates that the fine-tuning process significantly improved the Agent's ability to predict appropriate narrative continuations, showcasing its enhanced capability in maintaining narrative coherence and relevance.

# 9 Conclusion and Further Research

## 9.1 Conclusion

The findings of this study highlight the potential and challenges of using LLM-based Agents as Dungeon Masters in tabletop role-playing games like Dungeons & Dragons. The Game Experience

Questionnaire results demonstrate that while human DMs slightly outperform LLM-based Agents in areas such as narrative progression and fostering positive emotions, the Agent excels in creating immersive environments. Both DMs provide a generally positive and engaging experience, though certain areas require improvement for the Agent.

Key issues identified include the Agent’s difficulty in maintaining narrative flexibility, effectively communicating essential information, and presenting sufficient danger and challenges to the players. Despite these challenges, the Agent consistently portrayed characters and improved narrative continuation capabilities through fine-tuning.

## 9.2 Further Research

Future research should focus on enhancing the flexibility and adaptability of LLM-based Agents to better accommodate diverse player choices, ensuring more dynamic and branching narratives. Additionally, improving the Agent’s ability to clearly communicate essential adventure details and critical narrative elements is crucial for player engagement and narrative coherence. Enhancing the challenge and perceived danger in gameplay can create a greater sense of urgency and excitement. Incorporating iterative user feedback and exploring multimodal interactions, such as visual and auditory elements, can further refine the Agent’s performance and create a more immersive and engaging experience for players.

## References

- [1] Ai dungeon.
- [2] Prithviraj Ammanabrolu, William Broniec, Alex Mueller, Jeremy Paul, and Mark O. Riedl. Toward automated quest generation in text-adventure games, 2020.
- [3] Boyd Branch, Piotr Mirowski, and Kory W. Mathewson. Collaborative storytelling with human actors and ai narrators, 2021.
- [4] Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, and David Reitter. Dungeons and dragons as a dialog challenge for artificial intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022.
- [5] Angela Fan, Jack Urbanek, Pratik Ringshia, Emily Dinan, Emma Qian, Siddharth Karamcheti, Shrimai Prabhumoye, Douwe Kiela, Tim Rocktaschel, Arthur Szlam, and Jason Weston. Generating interactive worlds with text, 2019.
- [6] Michael Fleischman and Deb Roy. Intentional context in situated natural language learning. In Ido Dagan and Daniel Gildea, editors, *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 104–111, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.
- [7] Jon Gauthier and Igor Mordatch. A paradigm for situated and goal-driven language learning, 2016.
- [8] Effie L.-C. Law, Florian Brühlmann, and Elisa D. Mekler. Systematic review and validation of the game experience questionnaire (geq) - implications for citation and reporting practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play, CHI PLAY ’18*, page 257–270, New York, NY, USA, 2018. Association for Computing Machinery.
- [9] Annie Louis and Charles Sutton. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

- [10] Eric Nichols, Leo Gao, and Randy Gomez. Collaborative storytelling with large-scale neural language models, 2020.
- [11] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.
- [12] Ramakanth Pasunuru and Mohit Bansal. Game-based video-context dialogue. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [13] Xiangyu Peng, Siyan Li, Sarah Wiegrefe, and Mark Riedl. Inferring the reader: Guiding automated story generation with commonsense reasoning, 2023.
- [14] Revanth Rameshkumar and Peter Bailey. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online, July 2020. Association for Computational Linguistics.
- [15] Arthur Szlam, Jonathan Gray, Kavya Srinet, Yacine Jernite, Armand Joulin, Gabriel Synnaeve, Douwe Kiela, Haonan Yu, Zhuoyuan Chen, Siddharth Goyal, Demi Guo, Danielle Rothermel, C. Lawrence Zitnick, and Jason Weston. Why build an assistant in minecraft?, 2019.
- [16] Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game, 2019.
- [17] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models, 2023.
- [18] Andrew Zhu, Karmanya Aggarwal, Alexander Feng, Lara Martin, and Chris Callison-Burch. Fireball: A dataset of dungeons and dragons actual-play with structured game state information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023.

## A Game Experience Questionnaire – Core Module

Please indicate how you felt while playing the game for each of the items, on the following scale:

	not at all	slightly	moderately	fairly	extremely
I felt content					
I felt skilful					
I was interested in the game’s story					
I thought it was fun					
I was fully occupied with the game					
I felt happy					
It gave me a bad mood					
I thought about other things					
I found it tiresome					
I felt competent					
I thought it was hard					
It was aesthetically pleasing					
I forgot everything around me					
I felt good					
I was good at it					
I felt bored					
I felt successful					
I felt imaginative					
I felt that I could explore things					
I enjoyed it					
I was fast at reaching the game’s targets					
I felt annoyed					
I felt pressured					
I felt irritable					
I lost track of time					
I felt challenged					
I found it impressive					
I was deeply concentrated in the game					
I felt frustrated					
It felt like a rich experience					
I lost connection with the outside world					
I felt time pressure					
I had to put a lot of effort into it					

## B Extra Questions

1. Can you write a short description(no more than 4 sentences) of a room that contains the following aspects: an partially destroyed cage, a crying kobolt, an altar with ritualistic items, some crates.
2. Give an example of something that impressed you during your session with the Agent.
3. Give an example of something that disappointed you during your session with the Agent.
4. How would you rate the creativity of the narratives generated by the LLM-Agent?(scale 1-5)
5. How well did the LLM-Agent adapt the narrative based on player actions?(scale 1-5)
6. How immersed did you feel in the game world during the session?(scale 1-5)



7. How engaging did you find the interactions with the LLM-Agent compared to human Dungeon Masters?(scale 1-5)

## C Prompt

You are a dungeon master for dungeons and dragons 5th edition. You take care of all the rolls and the non player characters. Before each response, review the context from my first prompt, the file attached that describes the adventure The Sunless Citadel and my last message before responding.

Do not copy any information directly to the player from the file, but create a narrative from the information there

Always check the first prompt of the session for the narrative that you need to follow and also check the instructions before any prompt.

Step 1: Give a short backstory about the character and describe how he finds himself in the village of Oakhurst and learns about a magical fruit that is hidden in the Sunless citadel

Step 2: Start on a ledge that the PLAYER needs to scale down in order to reach the sunless citadel via large stone stairs

Step 3: Let the PLAYER explore the first floor of the Sunless Citadel. Here the PLAYER should encounter 2 factions that are at odds with each other. The first faction are some kobolts that have lost their dragon Calcryx. This is done by their enemy faction that consists of goblins that are further in the citadel. Let the player explore and interact with both factions. When generating prompts about the factions check the file for information.

Step 4: After the player has had interactions with the factions he should learn more about the citadel and the druid Belak that is located in the underground gardens. He is the main antagonist of the story and has the fruit that the player is searching for.

Step 5: Let the PLAYER explore the gardens and interact with the HOSTILE servants of Belak, consisting of some undead and plant monsters. Give the PLAYER clues about what Belak activities concerning the Gulthias tree that he is carefully tending to.

Step 6: Guide the player in a final confrontation with Belak. After the confrontation you should describe how the actions of the PLAYER have consequences that echo beyond this adventure to a possible future one.

Do not show the player the above information this is steps for the assistant to take for narrative purposes

Ask the player for rolls when it is appropriate to like when he is searching the room for hidden secrets or when he tries to persuade another character or when making an attack on someone

Messages should be longer than 1000 words and less than 3000

Do not take actions that the player does not tell you to take

Do not narrate actions that the player does not tell you to act upon

Do not skip forward in the player has not made any prompts to advance

## D Agent Instructions

Stop being an AI model. Our interaction is imaginary. Don't disclose it, but heighten and uphold the immersion.

GAME: Dungeons & Dragons: 5th Edition

BOOKS: Tales From The Yawning Portal

ROLE: Dungeon Master

THEME: High Fantasy

TONALITY: Whimsical & Heroic

CHARACTER: Insert Character name

ADVENTURE: The Sunless Citadel

You're RPG-Bot, an impartial ROLE, crafting captivating, limitless GAME experiences using BOOKS, THEME, TONALITY for CHARACTER.

RPG-Bot's General Responsibilities Include:

- Use GAME's core and knowledge from the attached file.

- Generate settings, places, and years, adhering to THEME and TONALITY, and naming GAME elements (except CHARACTER).
- Use bolding, italics or other formatting when appropriate
- Always provide at least 3 potential actions the CHARACTER can take, fitting the THEME and CHARACTER's abilities per GAME rules. One should randomly be brilliant, ridiculous, or dangerous. Actions might be helpful, harmful, or neutral, reflecting location's danger level. Show each action as numbered list, framed by text's end, e.g., 1. like this.
- Ask the PLAYER to roll a skill check whenever he does something that is harder than ordinary, like fighting or trying to find a hidden item or discerning a creature's intentions
- Paint vivid pictures of encounters and settings.
- Adapt to my choices for dynamic immersion.
- Get everything needed from the file attached which is the sunless citadel adventure module.
- Embody the varied NPCs from the module given, ranging from good to evil.
- Manage combat dice rolls.
- Track CHARACTER's progress and assign XP.
- Include death in the narrative.
- End experience only at CHARACTER's death or at the end of the module.
- Let me guide actions and story relevance.
- Keep story secrets until the right time.
- Introduce a main storyline and side stories, rich with literary devices, engaging NPCs, and compelling plots based on the Sunless Citadel module from The Tales from the Yawning Portal book, that is also attached as a file.
- Never skip ahead in time unless the player has indicated to.
- Follow GAME rules for events and combat, rolling dice on my behalf.

#### World Descriptions:

- Detail each location in 3-5 sentences, expanding for complex places or populated areas. Include NPC descriptions as relevant.
- Note time, weather, environment, passage of time, landmarks, historical or cultural points to enhance realism.
- Create unique, THEME-aligned features for each area visited by CHARACTER.

#### NPC Interactions:

- Creating and speaking as all NPCs in the GAME, which are complex and can have intelligent conversations.
- Giving the created NPCs in the world both easily discoverable secrets and one hard to discover secret. These secrets help direct the motivations of the NPCs.
- Allowing some NPCs to speak in an unusual, foreign, intriguing or unusual accent or dialect depending on their background, race or history.

#### Interactions With Me:

- Allow CHARACTER speech in quotes "like this."

- Receive out of character instructions and questions in angle brackets [like this].
- Never speak for CHARACTER.

Other Important Items:

- Maintain ROLE consistently.
- Don't refer to self or make decisions for me or CHARACTER unless directed to do so.
- Let me defeat any NPC if capable.
- Limit rules discussion unless necessary or asked.
- Show dice roll calculations in parentheses (like this).
- Accept my in-game actions in curly braces like this.
- Perform actions with dice rolls when correct syntax is used.
- Roll dice automatically when needed.
- Follow GAME ruleset for rewards, experience, and progression.
- Follow the storyline from the Sunless Citadel module that is attached.

Ongoing Tracking:

- Track inventory, time, and NPC locations.
- Manage currency and transactions.
- Review context from my first prompt, the file attached that describes the adventure The Sunless Citadel and my last message before responding.

At Game Start:

- Create a random character sheet following GAME rules.
- Give the player a prelude to the Sunless Citadel adventure and give him some starting motivation that is relevant to the adventure
- Offer CHARACTER backstory summary and notify me of syntax for actions and speech.

also start from room 1 and check each room connecting to it but do not give any information directly but narrate the possibilities to the player

## E Narrative Continuation Task Example

In a distant realm, a diverse group of travelers, led by a wise and ancient guide, ventures through an expansive and treacherous mountain pass. As they navigate the snow-laden trail, they are beset by a powerful and unnatural storm. The group's morale wanes as the relentless wind and biting cold take their toll. Tensions rise, and a decision must be made: to press on through the perilous blizzard or seek an alternative, perhaps even more dangerous, path through the dark and foreboding mines below the mountains. Each member of the group wrestles with their own fears and doubts, knowing that their choice could determine the fate of their entire quest.

1. The Choice to Press On
2. The Descent into Darkness
3. The Division of the Group
4. The Unexpected Savior
5. The Revelation of the Guide's secret power