CALYPSO: LLMs as Dungeon Masters' Assistants

Andrew Zhu[1], **Lara Martin**[2*], **Andrew Head**[1], **Chris Callison-Burch**[1]

[1]University of Pennsylvania
[2]University of Maryland, Baltimore County
{andrz, head, ccb}@seas.upenn.edu, laramar@umbc.edu

Abstract

The role of a Dungeon Master, or DM, in the game Dungeons & Dragons is to perform multiple tasks simultaneously. The DM must digest information about the game setting and monsters, synthesize scenes to present to other players, and respond to the players' interactions with the scene. Doing all of these tasks while maintaining consistency within the narrative and story world is no small feat of human cognition, making the task tiring and unapproachable to new players. Large language models (LLMs) like GPT-3 and ChatGPT have shown remarkable abilities to generate coherent natural language text. In this paper, we conduct a formative evaluation with DMs to establish the use cases of LLMs in D&D and tabletop gaming generally. We introduce CALYPSO, a system of LLM-powered interfaces that support DMs with information and inspiration specific to their own scenario. CALYPSO distills game context into bite-sized prose and helps brainstorm ideas without distracting the DM from the game. When given access to CALYPSO, DMs reported that it generated high-fidelity text suitable for direct presentation to players, and low-fidelity ideas that the DM could develop further while maintaining their creative agency. We see CALYPSO as exemplifying a paradigm of AI-augmented tools that provide synchronous creative assistance within established game worlds, and tabletop gaming more broadly.

Introduction

Dungeons & Dragons (D&D) (Gygax and Arneson 1974) is a tabletop role-playing game (TTRPG)—a collaborative storytelling game where a group of players each create and play as their own character, exploring a world created by and challenges set by another player known as the Dungeon Master (DM). It is the DM's role to play the non-player characters and monsters, and to write the overarching plot of the game.

As a co-creative storytelling game, Dungeons & Dragons presents multiple unique challenges for AI systems aiming to interact with it intelligently. Over the course of a game, which is played out across multiple sessions spanning a long duration of time (often multiple months to years), the DM and the other players work together to produce a narrative grounded in commonsense reasoning and thematic consistency (Ammanabrolu et al. 2020; Bergström 2011). As the group plays for longer, the players define more of the world and ad-hoc rules for interacting with it (van Velsen, Williams, and Verhulsdonck 2009). In order to make in-character decisions, each individual player must maintain a personal understanding of the game world which they build from the game history (Martin, Sood, and Riedl 2018) while keeping track of what information other players and their characters know (Zhou et al. 2023).

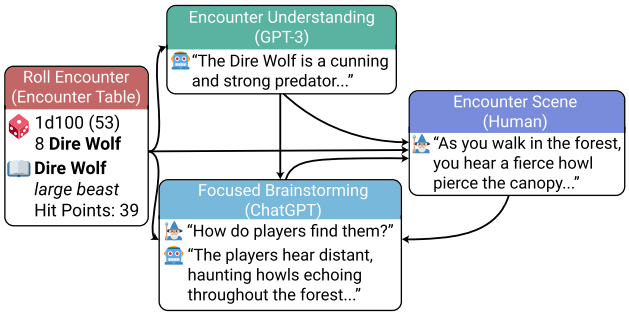

Figure 1: After rolling a random encounter (red), DMs can use LLMs with CALYPSO to help generate an encounter scene and digest information about monsters. CALYPSO can present monster information concisely (green) and brainstorm conversationally (purple) to help build a compelling narrative to present to players (purple).

By using an AI co-DM tool, human DMs can devote more mental energy to cognitively demanding tasks of being a DM, such as improvising dialog of NPCs (non-player characters) or repairing the script of their planned campaign. Furthermore, an AI co-DM would drastically reduce the barrier of entry into DMing. Therefore, an AI co-DM tool would be invaluable to the D&D community.

An effective AI co-DM tool should not only produce coherent and compelling natural language output for a DM to effectively use for inspiration but also account for an immense amount of background context and requirements for internal consistency—both within D&D rules and within a given scenario or campaign. Large language models (LLMs), such as GPT-3 (Brown et al. 2020) and ChatGPT (OpenAI 2022), have shown impressive abilities to generate

*Work done while at the University of Pennsylvania.

coherent text. Some (Callison-Burch et al. 2022; Zhu et al. 2023) have even applied LLMs to the problem of D&D dialog and narrative by finetuning the models with structured information. Whereas these works used structured information scraped from user data to fine-tune a single model, we use existing data in D&D source books to improve generation using zero-shot prompting with multiple models.

In this paper, we present a study in which we created a LLM-augmented tool to assist DMs in playing D&D. We employed the following methods:

1. We interviewed DMs to understand how they digest game information and learn design motivations for AI assistants in the domain.
2. We created a gameplay setting that allowed us to study D&D gameplay on a larger scale than other recent works and invited 71 players to participate.
3. We created a system of three LLM-powered interfaces, which we call CALYPSO (**C**ollaborative **A**ssistant for **L**ore and **Y**ielding **P**lot **S**ynthesis **O**bjectives), that DMs and players could use as they played D&D, and studied the ways in which DMs and players incorporated them into their creative process over four months using established HCI methods.

We show that language models are capable "co-DMs" – not a player in the same way that the human players and DM are, but still a synchronous agent that acts as a guide for the human DM. We provide insights into how TTRPG players actually want to use these tools and present validated solutions that can extend beyond the D&D domain. Our study shows that a system designed with these motivations in mind saw consistent prolonged usage among a community of creative writers.

Background and Related Work

Dungeons & Dragons in the Time of COVID

Traditionally, Dungeons & Dragons is played in person. Players use physical character sheets and monster stats referenced from books containing hundreds of prewritten "stat blocks" (as pictured in Figure 2a) (Perkins et al. 2014). DMs have the option to create a world of their own to play in (also sometimes called "homebrewing" a setting) or to set their game in a professionally written "module": a book containing a detailed outline of an adventure, including the setting, non-player characters, predesigned challenges and monster encounters, and lore. Previous works have explored methods of how to present information in these existing settings more clearly to DMs, such as through a computer-generated adventure flowchart (Acharya, Mateas, and Wardrip-Fruin 2021) or recommender systems for relevant entities in a scene (Perez, Eisemann, and Bidarra 2021).

Since the beginning of the COVID-19 pandemic, there has been a shift towards playing D&D online (Yuan et al. 2021). Rather than using physical character sheets and reference books while playing in person, a large number of groups instead play virtually using tools like D&D Beyond (Wizards of the Coast, LLC 2017) for virtual character sheets and reference books, Discord for messaging, virtual tabletops like Foundry (Foundry Gaming, LLC 2019) to simulate maps, and game state trackers like Avrae (Zhu and Wizards of the Coast, LLC 2016) to track character and monster stats. For inspiration and immersion, DMs also use online tools like dScryb (dScryb Inc. 2020), which provides prewritten text, Tabletop Audio (Roven 2014), which provides soundboards and soundscapes, and random tables published in D&D source books (Crawford, Perkins, and Wyatt 2014), which provide a prewritten set of options, for specific scenarios (e.g. encountering a dragon).

Large Language Models and D&D

Large language models (LLMs) are a recent development in the area of Natural Language Processing that have demonstrated emergent capabilities of understanding users' input and replying directly in the user's language (c.f. a machine language). A neural architecture based on the Transformer (Vaswani et al. 2017), they are capable of learning user-defined tasks with no additional training ("few-shot" or "in-context" learning) and referencing concepts defined in their large training corpus (Brown et al. 2020).

Although there has been some work looking at playing Dungeons & Dragons using earlier neural language models (Louis and Sutton 2018; Martin, Sood, and Riedl 2018; Rameshkumar and Bailey 2020), the introduction of LLMs has created a renewed interest in researching tabletop gaming. Callison-Burch et al. (2022) frame D&D as a dialogue challenge and examine whether LLMs are capable of predicting a player's next utterance based on the conversational history, finding that local game-specific state context is important for grounded narrative generation. Newman and Liu (2022) use LLMs to generate novel material (namely spells) that is consistent with the style and rules of the game. Zhou et al. (2023) create a system that models the intents of D&D players using LLMs to inform a surrogate Theory of Mind. Zhu et al. (2023) instrument a game state tracker to provide concrete actor stats and combat state, finding that LLMs are capable of producing interesting roleplay in combat scenarios and predicting the action a player will take. They highlight the importance of player and DM agency in LLM-generated texts, proposing that LLMs are better suited for assistant-style use cases. Kelly, Mateas, and Wardrip-Fruin (2023) present a preliminary work using LLMs to identify player questions from live transcriptions of gameplay and suggest in-character responses.

Santiago et al. (2023) have proposed multiple scenarios where LLMs and other generative AI models may be used to assist DMs, and discuss the ways AI may be used. In this workshop paper, they hypothesize the potential for AI to help inspire and take cognitive burden off the DM and provide brainstorming inspiration, but also weaknesses where AI may fall back onto overused tropes or underrepresent minority groups. In this work, we explore and expand upon many of these hypotheses through interviews with DMs. We create a system where DMs can fluently incorporate a LLM into their creative process and run a broad study on its use and failure cases.

LLMs have been explored as a writing assistant in other

modalities as well, using various methods to assist in collaboratively building a narrative. These works have examined the use of conversational agents (Coenen et al. 2021; Ippolito et al. 2022), writing in established settings (Akoury et al. 2020), and other human-in-the-loop methods (Chung et al. 2022; Roemmele and Gordon 2015; Samuel, Mateas, and Wardrip-Fruin 2016; Calderwood et al. 2020; Yang et al. 2022; Kreminski et al. 2022). There has also been work proposing LLMs for multimodal co-creative frameworks (Lin, Agarwal, and Riedl 2022). Overall, these techniques differ from D&D and other TTRPGs in that they primarily focus on a single writer/creator interacting with the system, rather than the multi-player experience in TTRPGs where all players directly interact with the story.

To our knowledge, our work is the first to examine concrete implementations of multiple unique interaction modalities in and outside of combat scenarios and the ways D&D players interact with language models on this scale.

Design Motivation

To better understand the friction DMs face in looking up reference material midgame, we conducted interviews and ran workshop sessions with seven DMs (referred to as D1-7 below) from a wide range of backgrounds before creating our system. Participants ranged from 1 to 39 years of experience playing D&D (various editions). In these sessions, we asked DMs how they approached improvising encounters – i.e., to run random encounters that are generated on the fly (usually by rolling on an encounter table). In random encounters, DMs do not have time to research the monster's stats and lore beforehand and think of backstories as to why the monster ended up in a particular setting. From these interviews, we identify several ways how an AI system could be helpful to DMs:

Inspiration. As proposed by Santiago et al. (2023), we find that DMs desired the ability to use a language model to generate the first draft of an encounter, which they could then build on top of with their own ideas (D1-3). Different DMs envisioned giving the system varying amounts of control over the narrative. D3 expressed that they would want a system to write a scene that they would then vet and choose whether to present it verbatim to their players, edit it to their liking, or use as inspiration to overcome writer's block. D1 and D2 envisioned using the system's generation verbatim to present an initial scene to players while they either read the complete text of the monster description (D2) or to reduce cognitive load (D1).

Strategic Copilot. One DM mentioned that managing both narrative gameplay and tracking monster stats and mechanics overwhelmed their short-term memory, and expressed interest in a system that could aid them in making strategic decisions and acting as a high-level copilot. They expressed that the large amount of low-level management was a barrier to them running more D&D, and that they wanted to "feel more like an orchestra conductor over someone who's both putting down the train tracks AND fueling the train" (D4).

Another DM said that DMs often fail to take into account monsters' unique abilities and stats when running encounters, making simplifications to manage a large number of monsters. For example, a monster with very high intelligence and low dexterity attempting to move sneakily "should know not to move and make a bunch of noise" (D6).

Thematic Commonsense. We asked DMs what parts of monsters' game statistics they found to be the most important for their understanding of how to use a monster in their game, and found that multiple DMs used a concept of "baseline" monsters to gain a broad understanding of a monster when they first encounter it. The idea of the baseline monster was not to find a specific monster to compare another to, but to determine which parts of an individual monster's game statistics to focus on, and which parts to use prior thematic commonsense to fill in.

In this context, we define *thematic commonsense* as the DM's intuitive understanding of D&D as a game with medieval fantasy themes, and how they might draw inspiration from other works of fantasy literature. For example, a DM might intuitively understand that a dragon is a kind of winged reptile with a fire breath based on their consumption of other fantasy works, reason that all dragons are capable of flight, and focus on a particular dragon's unique abilities rather than flight speed (D7). Although D&D reference material does not include an explicit description of the dragon's fire breath, the DM might base their narration on depictions of fire breath from other authors.

We find this similar to the idea of a *genus-differentia* definition (Parry and Hacker 1991), in that DMs use their general background understanding of fantasy settings to define their personal *genus* and supplement prior knowledge by skimming monster reference books for *differentia*. This suggests that co-DM systems should focus on helping DMs extract these *differentiae*, and that they also require the same extensive background knowledge as the user. For the D&D domain, we believe that LLMs such as GPT-3 (Brown et al. 2020) have included sufficient information on the game and the game books themselves in their training corpus so as to establish such a background knowledge. However, we are interested in methods for establishing this thematic commonsense knowledge for works not included in models' training data in future work.

Simple Language. Multiple DMs emphasized that they would like a co-DM system to present monster information in plain language, rather than the elaborate prose found in game reference manuals (D3-6). As a work of fantasy literature, D&D publications (including reference manuals) often use heavy figurative language and obscure words. For example, the first paragraph of an owlbear's description reads:

> An owlbear's screech echoes through dark valleys and benighted forests, piercing the quiet night to announce the death of its prey. Feathers cover the thick, shaggy coat of its bearlike body, and the limpid pupils of its great round eyes stare furiously from its owlish head (Crawford, Mearls, and Perkins 2018, pg. 147).

This style of description continues for seven additional

paragraphs. On average, across all D&D monsters published on D&D Beyond, a monster's description and list of abilities contains 374 words (min: 0, max: 2,307). DMs often use multiple monsters together in the same encounter, compounding the amount of information they must hold in their mind.

Monster descriptions often include descriptions of the monster, its abilities, and lore. Some DMs' preferred method of referencing monster lore while running the game was to skim the full monster entry, and the complex and long prose often led to DMs feeling overwhelmed (D4, D5). Other DMs wanted a short and salient mechanical (i.e. focusing on monster's game abilities and actions) description, rather than a narrative (lore and history-focused) one (D3, D6).

Overall, the complexity of monster descriptions led DMs to forget parts of monsters' lore or abilities during gameplay (D5) or use overly broad simplifications that did not capture an individual monster's uniqueness (D6). While offline resources exist to help DMs run monsters (e.g. Amman (2019)), they cannot account for the environment or generate a unique scenario for each encounter with the same monster. We believe that LLMs' capability to summarize and generate unique material is particularly applicable to these challenges.

Implementation

In this section, we describe the three interfaces we developed to provide DMs with the sorts of support they desired. These interfaces were designed with "in the wild" deployment in mind:

1. Encounter Understanding: a zero-shot method to generate a concise setup of an encounter, using GPT-3.
2. Focused Brainstorming: a conversational method for DMs to ask additional questions about an encounter or refine an encounter summary, using ChatGPT.
3. Open-Domain Chat Baseline: a conversational interface without the focus of an encounter, using ChatGPT.

Our implementation differs from other efforts to develop AI-powered co-creative agents in two ways. First, compared to models where the AI acts as the writer, AI-generated content is not necessarily directly exposed to the audience. CALYPSO only presents ideas to a human DM, who has final say over what is presented to the players. Second, compared to co-writing assistants where the writer has plentiful time to iterate, the time between idea and presentation is very short. Since the DM uses CALYPSO in the midst of running a real game, CALYPSO should be frictionless to adopt and should not slow down the game.

Encounter Understanding

The first interface we provided to DMs was a button to use a large language model to distill down game statistics and lore available in published monster stat blocks. To accomplish this, we prompted GPT-3 (Brown et al. 2020) (specifically, the text-davinci-003 model) with the text of the chosen encounter, the description of the setting the encounter was taking place in, and the game statistics and lore of each monster involved in the encounter. The full prompts are available in the appendix.

We began by presenting the LLM with the task to summarize monsters' abilities and lore and the environment. We collected feedback from DMs after generating the extracted information by allowing them to select a positive or negative feedback button, and optionally leave comments in an in-app modal. This interaction is illustrated in Figure 2.

Summarization. At first, we prompted GPT-3 to "summarize the following D&D setting and monsters for a DM's notes without mentioning game stats," then pasted verbatim the text description of the setting and monster information. For decoding, we used a temperature of 0.9, top-p of 0.95, and frequency and presence penalties of 1. Based on feedback from DMs (discussed below), we later changed to a more abstract "understanding" task described below.

Abstractive Understanding. In the understanding task, we prompted GPT-3 with the more abstract task to help the DM "understand" the encounter, along with explicit instructions to focus on the unique aspects of each creature, use information from mythology and common sense, and to mention how multiple creatures interact with each other. After these instructions, we included the same information as the *Summarization* task above. Finally, if a monster had no written description, we included instructions in place of the monster's description telling CALYPSO to provide the DM information from mythology and common sense. For decoding, we used a temperature of 0.8, top-p of 0.95, and a frequency penalty of 0.5.

Focused Brainstorming

To handle cases where a single round of information extraction was not sufficient or a DM had additional focused questions or ideas they wanted assistance elaborating, we also provided an interface to open a private thread for focused brainstorming. Available at any time after an encounter was randomly chosen, we provided the same encounter information as in the *Encounter Understanding* interface as an initial prompt to ChatGPT (i.e., gpt-3.5-turbo) (OpenAI 2022). If the DM had used the *Encounter Understanding* interface to generate an information block, we also provided it as context (Figure 4). The full prompts are available in the appendix. For decoding, we used a temperature of 1, top-p of 0.95, and a frequency penalty of 0.3.

Open-Domain Chat Baseline

Finally, we made a baseline open-domain chat interface available to all players, without the focus of an encounter. As this interface was available at any time and open-ended, it helped provide a baseline for how DMs would use AI chatbots generally. To access the interface, users were able to run a bot command, which would start a new thread. We prompted ChatGPT to take on the persona of a fantasy creature knowledgeable about D&D, and generated replies to every message sent in a thread opened in this manner. For decoding, we used a temperature of 1, top-p of 0.95, and a frequency penalty of 0.3. Unlike the private threads created by the *Focused Brainstorming* interface, open-domain