FIREBALL: A Dataset of *Dungeons and Dragons* Actual-Play with Structured Game State Information

Andrew Zhu[1]**, Karmanya Aggarwal**[1]**, Alexander Feng**[1]**,**
Lara J. Martin[2*]**, and Chris Callison-Burch**[1]
[1]University of Pennsylvania
[2]University of Maryland, Baltimore County
{andrz, karmanya, ahfeng, ccb}@seas.upenn.edu, laramar@umbc.edu

Abstract

Dungeons & Dragons (D&D) is a tabletop role-playing game with complex natural language interactions between players and hidden state information. Recent work has shown that large language models (LLMs) that have access to state information can generate higher quality game turns than LLMs that use dialog history alone. However, previous work used game state information that was heuristically created and was not a true gold standard game state. We present FIREBALL, a large dataset containing nearly 25,000 unique sessions from real D&D gameplay on Discord with true game state info. We recorded game play sessions of players who used the *Avrae* bot, which was developed to aid people in playing D&D online, capturing language, game commands and underlying game state information. We demonstrate that FIRE-BALL can improve natural language generation (NLG) by using Avrae state information, improving both automated metrics and human judgments of quality. Additionally, we show that LLMs can generate executable Avrae commands, particularly after finetuning.

1 Introduction

Dungeons & Dragons (D&D) (Gygax and Arneson, 1974) is a tabletop roleplaying game in which players assume the roles of characters in a fantasy adventure. Play is conducted primarily through natural language, with players roleplaying as their characters and describing their actions. Meanwhile another player, the Dungeon Master (DM), controls the fictional story world: setting obstacles, goals, and adventures, controlling monsters, and interpreting players' actions in the context of the rules of the game. Although the DM makes a lot of the final decisions, the game is ultimately a collaborative storytelling experience.

Due to its use of natural language as actions, each individual player must maintain a personal

---
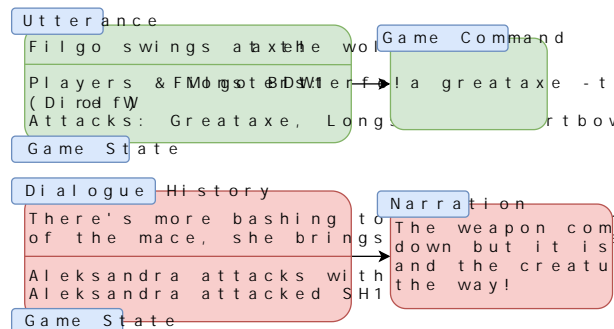
Figure 1: Examples of our Utterance to Command task (top), which takes in an utterance and a game state to produce an Avrae command, and State to Narration task (bottom), which produces a narration given a dialogue history and game state information.

understanding of the game world which they build from conversational history and using Theory of Mind (Martin et al., 2018; Zhou et al., 2022). The natural language action space also means that the AI needs the ability to adequately perform tasks such as language generation, language understanding, and planning (Callison-Burch et al., 2022).

Although AI's capabilities in this space are still nascent, Callison-Burch et al. (2022) have shown that D&D dialog can be improved by adding state information into the input of a large language model (LLM). However, the state information presented in that work was heuristically created using regular expressions and machine learning classifiers. Thus it cannot be considered true ground truth state information. Our work is unique because it provides true ground truth state information.

We use this data for two tasks: *Utterance to Command* and *State to Narration*. In the first task, a model is given a game state and turn of the game (roleplay in natural language), and must predict the corresponding command that matches the *intent* of the roleplay. The second task is a constrained creative natural language generation task: given a state change resulting from a command execu-

tion, generate a narration that describes the results. Figure 1 demonstrates both tasks.

Our contributions are as follows:

- We present FIREBALL—a dataset of over 8M gameplay utterances, 2.1M commands, 1.2M gameplay states, and 160K unique actors (player & non-player characters)[1]. This is the first dataset of this size that includes detailed game state and character information for each scenario.
- We show that large language models such as GPT-3, can extract relevant information from natural language in order to produce commands that are capable of being run by the game environment.
- We demonstrate that LLMs, when finetuned on this dataset, can generate more grounded narrative text compared to language models tuned without game state information.

By incorporating structured state information into language understanding and NLG tasks, we hope to help pave the way toward more ambitious creative generation goals, such as consistent long-form narrative generation and video games that can convert language input into discrete game actions or generate narrations and dialogues based on the game state.

2 Related Work

Previous papers have outlined the challenges of Dungeons & Dragons as an AI problem and examined various aspects of the game (Ellis and Hendler, 2017; Martin et al., 2018). Subsequently, a number of datasets in the D&D space have been created (Louis and Sutton, 2018; Rameshkumar and Bailey, 2020; Si et al., 2021; Callison-Burch et al., 2022; Papazov et al., 2022), but these datasets either do not include game state information or include only an inexact game state, which lacks grounded, verified attributes, such as those included in our dataset. Others have looked at using AI for subsets of D&D gameplay, such as generating spell descriptions (Newman and Liu, 2022) or simulating combat (Gama Vila Nova et al., 2019).

In addition to the gold-labelled game state, we include all of the attributes from the following papers within our FIREBALL dataset. Louis and Sutton (2018) present a dataset pairing independent roleplay utterances with narrative descriptions of the
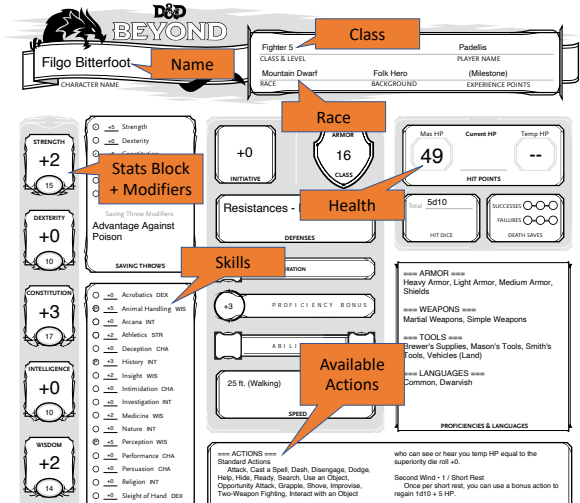

Figure 2: A snippet of a Character Sheet for the character Filgo Bitterfoot. Players use pages like this to keep track of information about their D&D characters. We capture all this information in our state, in addition to information related to combat. Visit our repo (https://github.com/zhudotexe/FIREBALL) for the full document.

characters that said them, and show that these descriptions influence the style of their utterances. More recently, Callison-Burch et al. (2022) frame D&D as a dialogue challenge, approaching the challenge of generating cohesive and interesting gameplay utterances given the conversational history and a predicted game state. Papazov et al. (2022) present a pilot study on understanding player intent and translating it to a high level game action, represented as an Avrae command. Their workshop paper examines similar tasks, but studies only 100 hand-annotated events.

Outside of the D&D domain, the task of translating natural language into a bash command (Lin et al., 2018) or regex (Locascio et al., 2016) is similar to our Utterance to Command task, although these languages do not require conditioning on a context. For another natural language-centric game, Diplomacy, Bakhtin et al. (2022) demonstrate the importance of providing game state information to language models.

3 Playing D&D Using Avrae

In Dungeons & Dragons, a group of players each create and play as a character. Characters have classes (such as wizard or barbarian), fantasy races (such as Elves, Gnomes, and Dragonborn), hit points (denoting their health), statistics that govern how well they can do certain actions, and invento-

---

[1] Our FIREBALL dataset is available here: https://github.com/zhudotexe/FIREBALL

ries of items (armor, weapons, potions, etc). These game state elements are stored on a character sheet (Figure 2). One player takes on the role of the Dungeon Master (DM). This special player creates the world in which the story is told, role plays all of the characters the other players interact with, and acts as arbiter of the rules of the game.

There are two main modes of gameplay –in-combat and out-of-combat– which have different styles of play. In-combat play simulates battles between characters and monsters, and involves turn taking and tracking stats. Out-of-combat play is characterized by freeform collaborative storytelling. Both elements of the game involves rolling dice to determine the success of players' actions (like attempting to attack a monster). The die's outcome is then modified with a stat that represents the character's skill in performing that particular action (e.g., +3 for acrobatics). If the dice roll plus the modifier was above a threshold that the rules or the DM determines (called a difficulty class, or DC), then the action succeeds. Otherwise it fails. The DM then narrates the results.

For example, if an attack hits a monster, the DM might narrate the player's blow and how the monster reacts, taking into account information like the attack's damage type and how many hits the monster has taken previously.

Traditionally, D&D is played in person with characters' stats written out on physical character sheets and monster stats referenced from books containing hundreds of prewritten "stat blocks". To track a stat that changes frequently like hit points, players and DMs use paper and pencil or whiteboards, performing math in their head and writing over the previous value when it changes. Some players also use maps and miniatures to track where characters and monsters are located relative to one another and to aid immersion in the game world. Since the beginning of the pandemic, a large number of groups have moved online using tools like Discord (a messaging program), virtual tabletops that simulate maps, and game state trackers like Avrae rather than physical mediums.

Avrae is a Discord bot that was designed to help people play D&D online. It allows players to import their character sheets, allows DMs to access a database of monsters, and simulates dice rolls. During combat, Avrae tracks the game state. This state contains detailed information including the list of participants in the battle, their stat blocks,

their current HIT points, and their available actions. Avrae allows players to execute commands representing their characters' actions. It performs a simulated dice roll, adds the player's modifiers, and determines the success or failure of the roll. Avrae then updates the game state, adjusting things like hit points, and turn tracking.

A simplified example of interacting with Avrae might look like the following example:

> **Player**: Filgo crouches down in the bush, loosing an arrow at the dire wolf charging towards him.
>
> **Player**: `!attack longbow -t DW1`
>
> **Avrae**: (Rolls dice and displays the results of the attack and damage dealt, including the new health of the dire wolf.)
>
> **Dungeon Master**: Your arrow flies true and the beast lets out a shrill howl as it pierces its matted fur. It's low on health now, so on its turn it'll retreat.

In actual play, an average of 3-8 players (including the DM) take turns interacting with Avrae. By instrumenting Avrae to record these commands and messages before and after a user's inputted command, we collect a rich set of structured gameplay. Appendix B contains a full list of recorded events and their descriptions.

4 Dataset

We worked with the developer of Avrae to instrument it to collect game transcripts and log game state information. The data collection was approved by Wizards of the Coast, the game company that owns D&D and Avrae, as well as by our institution's IRB and the Bot Safety team at Discord. Any players who participated in our study provided their informed consent.

4.1 Data Collection

Participants were recruited from English-speaking "play-by-post" D&D Discord servers, where players and Dungeon Masters play by taking turns posting in a Discord text channel to describe their moves. To accomplish this, we made a website explaining the study, and recruited server admins to review the study and opt-in to participate by announcing our research study on the Avrae Discord server. When players began combat in an opted-in server, they were presented with a message informing them of the recording. Players could opt-out

individual combat instances, server admins could stop all recording on a server at any time, and we provided a public form for requests to delete data. For each actor (player or monster), we record a detailed state; Table 6 in the Appendix lists all available attributes and their potential relevance to NLG tasks. Similarly, recorded actions include the detailed results of each dice roll, such as whether a given attack hit its target or a spell succeeded (a list of all action attributes is available in Appendix C).

In the following sections, we refer to our data as triples consisting of a command and its corresponding state change, any relevant utterances before the command ("preceding" utterances), and any relevant utterances after the command ("following" utterances). Each of these commands corresponds to an action that an actor in combat can take, such as attacking with a weapon, casting a spell, or using a special ability.

4.2 Utterance-Action Alignment

To align utterances with their corresponding state changes, we match each utterance with its chronologically nearest state change, and tag all utterances that occur chronologically before their corresponding state change as an utterance motivating the command (the "preceding" utterances), and all utterances that occur chronologically after their corresponding state change as narration of the state change (the "following" utterances). These alignments create a prototypical triple as described above. Within each triple, we discard any utterance containing less than five words.

4.3 Authorship Filtering

Within each triple, we identify the user who issued the commands, and the Dungeon Master hosting the combat. We discard any utterances within each triple which are not authored by one of these users. Additionally, we discard any triple where the commands originate from multiple different actors, which may occur if a single user is controlling multiple different creatures in a group. Finally, we discard any triple which has neither any "preceding" utterances nor "following" utterances.

4.4 IC/OOC Classification

We further distill the set of "following" utterances by training GPT-3 Ada (Brown et al., 2020) to distinguish between "in-character" (IC) utterances

from "out-of-character" (OOC) utterances. In-character utterances are what the player says speaking as their character or to describe their character's actions. They might look like this:

> Filgo puts a hand on his axe, uneasy after the shaking he'd felt from the ground.
>
> "Is someone there?"

Meanwhile, out-of-character utterances occur across players/the DM when not speaking as any particular character. This dialog might be to discuss rules or strategy, or might be unrelated to gameplay entirely. Out-of-character utterances might look like this:

> How much health do you have left?
>
> I'll move back 30 feet after.
>
> BRB, going to the bathroom.

To distinguish between these categories, we fine-tuned a classifier that was pretrained on Giant in the Playground data (Callison-Burch et al., 2022), forum roleplay data which includes labels for in-character and out-of-character posts, on a hand-labelled set of 750 utterances randomly sampled from our dataset. The classifier achieved an accuracy of 94% on a validation set of 125 utterances. We then applied the classifier to each utterance in our dataset and discarded any out-of-character utterances from the "following" set since in-character text is usually more interesting and evocative. Finally, we also removed sections of utterances contained in parentheses, which usually indicate OOC speech, from the "following" set.

4.5 Dataset Size

Our dataset contains 25k unique combat scenarios, including 8M utterances from 3.6k unique authors covering 1.3M unique combat states. Table 1 contains a breakdown of the distribution of commands in our dataset, organized by command category.

5 Utterance to Command Task

Our first task aims to predict the game command that a player or Dungeon Master intended to use, given the utterances since their last turn in combat. To successfully predict a command from an utterance, a model must be able to predict the user's

---

[2]Checks and saving throws contained as part of an action are not included in this category.

Type	Invocations	Example
Combat	713,568	`!init next`
Actions	608,527	`!cast fireball`
Custom	313,898	`!map`
Character	97,033	`!game hp`
Checks[2]	95,413	`!check arcana`
Dice Rolls	76,990	`!roll 1d20`
Other	204,174	`!help`
Total	2,109,603	

Table 1: The number of command invocations in the dataset, organized by command category.

intent, which actors the user intended to target, and ground both these predictions in the game state. For example, in the scenario illustrated in Figure 1, a dwarf named Filgo is fighting a Dire Wolf. On his turn, his player narrates that Filgo attacks with his axe, then runs the command to target the monster with his attack. Notice how, in this example, the player references the target dire wolf by its creature type ("the wolf"), rather than its name in the game state ("DW1").

To accomplish this task, we provide the state information included in our dataset—namely the list of actors participating in combat and any information about those actors, such as their monster type and current hit points—to the models. The full prompt for the example mentioned above is available in Appendix F.

After our distillation passes, our dataset contains 120,000 aligned utterance-command pairs. We examine the accuracy of predicted commands on both a token level and by injecting predicted commands into the Avrae system. Finally, we also examine the performance of models without game state information included to demonstrate the importance of the game state.

5.1 Models

We use GPT-3 (Brown et al., 2020) Davinci models (as of Dec. 2022) as a base. Finetuned models are using standard Davinci, while few-shot models use Davinci-002. For the Utterance to Command generation task, we evaluate four main treatments,

- **FT + S:** The base model is finetuned on a sample of 30K examples from FIREBALL with state information presented in the prompt.
- **FT:** The base model is finetuned on a sample of 30K examples from FIREBALL without any state information presented in the prompt.
- **FS + S:** The base model is presented 3 exemplars (few-shot) sampled from FIREBALL with relevant state information.

- **FS:** The base model is presented 3 exemplars sampled from FIREBALL without any state information.

5.2 Evaluation

Each of the generation tasks is evaluated independently. Since the command generation task is more akin to a structured generation task, we evaluate only on objective correctness rather than subjective quality of generated text. We evaluate the generated commands over four quantitative metrics: pass rate, unit tests, RougeL (Lin, 2004) and Average Sentence Gleu (SGleu) (Mutton et al., 2007).

We first seek to evaluate whether the generated command is a valid Avrae command by simply passing the command to Avrae and checking for successful execution. Similar to Chen et al. (2021) we calculate a *pass rate* metric that determines the proportion of generations that constitute valid Avrae commands. To calculate this metric, we have each model generate commands for 1000 utterances randomly sampled from a held out test set, and simply count the proportion of generations that Avrae is able to successfully execute.

Second, we evaluate what proportion of generated commands would result in the desired state update through a number of hand-written *unit tests*. Each test accepts a predefined combat state, an utterance, and the corresponding model-generated command and validates assertions on the combat state update. We took 10 common scenarios seen in D&D for these unit tests, generating 10 commands for each scenario-model pair. Since these generated commands sometimes have repeats, we take the n unique commands from the set of generations and validate which proportion of the generated samples pass the handwritten unit tests by running assertions on the updated combat state after running the command through Avrae.

Lastly, we perform a qualitative analysis to better understand the nature of the grounding. We took two popular spells (Bardic Inspiration and Fireball) and hand constructed a scenario for each. We then perturb the prompts that we present to the model to study the model's sensitivity to inputs.

5.3 Results & Discussion

Table 2 displays the objective evaluations of the four treatments in the Utterance to Command generation task. We see that the model which was finetuned on FIREBALL with the state information

Model	Pass Rate	Unit Tests	SGleu	RougeL
FT+S	**0.726**	**0.65**	**0.355**	0.75
FT	0.235	0.234	0.189	0.551
FS+S	0.432	0.429	0.325	**0.771**
FS	0.319	0.25	0.246	0.598

Table 2: Aggregated Results from the four models on the Utterance to Command Generation Task. SGleu refers to the average Sentence Gleu.

Model	Perplexity	BERT	ROUGE
DIALOG	208.97	0.8458	0.1077
COMMAND	**156.98**	0.8421	0.0919
FIREBALL-SHORT	202.39	**0.8478**	0.1087
FIREBALL-FULL	208.98	0.8476	**0.1156**
Human	452.653	N/A	N/A

Table 3: Perplexity, BERTScore, and ROUGE-1 scores of our models and human-written responses.

(FT+S) significantly outperforms all other models (i.e., both models without the state information and the few-shot models). Within the perturbation results (detailed in Appendix D), we notice that the FT+S model can accurately gauge the state of the actors in combat. For example, when asked to cast a Fireball at injured enemies, it successfully parses the prompt to find the subset of enemies that were injured and only targets them in the resulting command. Further, the model can accurately determine which of the prepared spells correspond to an utterance. For example, an utterance that would have generated the Fireball spell generates the spell Burning Hands instead if Fireball is removed from the prepared spell list.

6 State to Narration Task

In this task, we want to generate a narrative utterance describing the effects of a player's actions, given all of the state changes since the start of the player's turn in combat.

For example, in the scenario illustrated in Figure 1, a party is fighting a Sea Hag. On the cleric's turn, she attacks the hag with her mace, but misses. After seeing the result of her action, she narrates the miss, referencing the result as the hag dodging the attack. The full prompt for this example is available in Appendix E.

After our distillation and filter passes (Sections 4.2-4.4), our dataset contains 43,000 aligned state-utterance pairs. To examine the importance of the game states provided in our dataset, we compare our results to methods that do not include the game state, such as dialog continuation (Callison-Burch et al., 2022) and predicting the narration given only the command that was run (Papazov et al., 2022).

6.1 Models

We finetuned four GPT-3 (Brown et al., 2020) models on different data to determine the effect of state and dialog history inputs on generation. Each

model uses Davinci (as of Dec. 2022) as a base model, using 20,000 state-utterance pairs.

- **DIALOG:** Our first baseline model. This model is only given the last 5 messages of chat history, and fine-tuned to predict the next utterance that continues the dialog. It is not given any information about the game.
- **COMMAND:** Our second baseline model. The model is only given the command that was run to take the player's action, and fine-tuned to predict the corresponding utterance.
- **FIREBALL-SHORT:** Similar to DIALOG, but also contains the mechanical description of the action's results.
- **FIREBALL-FULL:** All information given to FIREBALL-SHORT plus the full actor list, target list, and detailed attributes of the caster.

6.2 Automated Evaluation

For the combat State to Narration generation task, we leverage standard text generation metrics: perplexity using a GPT-2 model (Radford et al., 2019) as a baseline, BertScore (Zhang et al., 2019), and ROUGE (Lin, 2004). All metrics aside from perplexity are calculated using the human narration as a reference. The results of our automated evaluation are available in Table 3.

We note that automated metrics are not particularly suited for evaluation of creative natural language generation. Perplexity is a measure of how "unexpected" a sequence is to a language model, which does not directly correlate with the *quality* of creative generation. Furthermore, BertScore and ROUGE evaluate similarity to a reference, which is an imperfect fit for our task where two narrations can differ substantially yet both be of high quality. These limitations are evident in the disparity in results between automated and human evaluation, which is expected given previous work that reached

Model	Sense	Specific	Interest
DIALOG	0.36	0.27	4.27
COMMAND	0.41	0.37	4.72
FIREBALL-SHORT	0.52	**0.48**	**4.98**
FIREBALL-FULL	**0.55**	0.47	4.6
Human	0.54	0.48	4.91

Table 4: Average sense, specific, and interestingness scores of our models and human-written responses.

similar conclusions (Sagarkar et al., 2018; DeLucia et al., 2021).

6.3 Human Evaluation

We also perform a human evaluation to assess the quality of the generated utterances. In total, we recruited 45 evaluators from the Avrae user base based on their experience with Avrae and D&D. All evaluators had played D&D using Avrae before: 37 had used Avrae for over a year and 37 had been the Dungeon Master of a game using Avrae. Evaluators were rewarded with a set of codes for digital goods on D&D Beyond with a market value of $36 for completing the rating task.

We provided each evaluator a version of the context that was provided to the models: the last fifteen messages sent in a channel, the casting actor and their description, a list of actors in combat, and the current state of those actors. Along with each context, we provided one generated utterance from each model along with the true utterance sent by a human. The evaluators were asked to rate each output along three dimensions, following the evaluation procedure used for the Meena LM (Kulshreshtha et al., 2020) and D&D Dialogue dataset (Callison-Burch et al., 2022):

- Does the response make sense? (yes/no)
- Is the response specific? (yes/no)
- How interesting is the response? (10 point scale)

Each evaluator rated 3 to 7 scenarios randomly drawn from a set of 75, with at least 3-way redundancy for each scenario. The full annotator instructions and a mockup of the annotation interface are given in Appendix G.

6.4 Results & Discussion

The results of our human evaluation are tabulated in Table 4. Both FIREBALL models outperform the baseline models in sensibility and specificity by an average of 15 percentage points (significant to $p < 0.01$), and on average perform similarly to a human ($p > 0.5$). A detailed analysis of significance can be found in Appendix H. Generally, models that were aware of the game context (including COMMAND) were more interesting than the model tasked with simply continuing the chat history, and comparable to human performance.

It may seem unusual for the human performance to be so low, only making sense to an experienced D&D player about 50% of the time. One explanation could be that human-written responses were more likely to refer to background knowledge not provided in the model's context, and therefore may have caused raters to mark the response as nonsensical. We compiled qualitative feedback from our evaluators to provide some insight into why this may be, as well as identify some common failure cases of our models. We summarize some of the reoccurring themes here.

Removing Player Agency. The most common theme among evaluators' feedback was that they did not want the model to take control of their character away from the player. One evaluator noted that "*several [narrations] had player characters acting or speaking in response to the action. That's something I would never want a human DM doing unprompted and it might be frustrating to have the bot look like it's trying to control what my character does or says.*" This problem extended to the Dungeon Master's role, as well; one evaluator mentioned that some AI responses would specify creatures' movement and "drag [their] encounters down the hallway."

Player agency is an especially challenging aspect of the game to maintain while training language models with real player utterances, as the training data available naturally makes decisions for and speaks as the player. Multiple evaluators suggested that language models should be used to provide assistance and inspiration for players rather than directly acting on behalf of the player, allowing a human to make the final decision for their characters.

Hit Point Reasoning. Another common failure case was models' tendency to describe any damage to a target, regardless of the target's true remaining health, as a kill. According to one evaluator, "*I ran into a scenario where the text was absolutely great, but incredibly wrong. A player scored a critical hit on a Mind Flayer with a bow, and left the mind flayer bloodied. The long text stated the scenario beautifully, even got the weapon right... but said it*

Last Human Utterance	"Dogs... so many of them making it more annoying then anything." Umbrage didn't want to waste anything big. So he pulled out his pipes, taking the attack of opportunity as he would start to play. "Fear me!"
Command Run	`!i cast "Cause Fear" -t dd1 -t dd3 -t dd4 [...]`

Model	Narration
DIALOG	The music was certainly not pleasant, but it seemed to have an effect on the dogs. They became panicked for a moment, but were able to maintain their aggression. It was just enough for the party to act.
COMMAND	Cuco proceeds to try and scare the ducks. So far it wasn't working.
FIREBALL-SHORT	As the first dog fell, the others seemed to flee once more in pure terror of the seemingly unending tunes of the pipes.
FIREBALL-FULL	The dogs were in the middle of a panic. But still they fought on!
Human (Original Narration)	Moving back into the melee with the one he was originally fighting. As he drew ready.

Table 5: Examples of predicted completions from each model given the last human utterance, the command that was run, or both. In this scenario, a Dragonborn Bard named Umbrage has just cast the Cause Fear spell against a pack of Death Dogs. As a result of the command being run, three of the dogs fail to save and are afraid; four resist the effect.

killed the Mind Flayer."

We noticed this failure case was more prevalent among the DIALOG and COMMAND models, demonstrating that our state-aware models were able to appropriately ground the generated narrations.Nevertheless, even the state-aware models often made this mistake. We have seen in our dataset that humans tend to only narrate critical portions of a fight, like when an enemy is killed. This reporting bias creates a disproportionately large number of narrations in our training set that reference a creature being killed. Therefore, the models trained on this data skew towards narrating that monsters are killed.

Table 5 shows a scenario that exemplifies many of these aspects. Without the guidance of previous dialog turns, the COMMAND model hallucinates names for the caster and the targeted creatures. The DIALOG, FIREBALL-SHORT, and FIREBALL-FULL models produce plausible next utterances based on the details provided in the player's utterance but exhibit some of the discussed failure cases: it acts as the Dungeon Master to narrate that the dogs have fled and it references the first dog dying, which it did not (in the game state, the dog still has a full 39/39 hit points). For reference, the true next utterance as written by the player does not mention any effect of the spell, instead focusing on the character's movement. The full game state and chat history associated with this example is included in Appendix I.

7 Conclusions

We have demonstrated how the FIREBALL dataset can be used to predict game commands that correctly match a player's intent and generate cohesive and grounded narration from the results of a game action. Our Utterance to Command model is capable of translating roleplay into game-specific actions and can aid novice users, reducing the amount of time players spend looking up documentation and allowing them to play the game more. Our State to Narration model can help inspire the Dungeon Master and take some of the cognitive load off of repetitive writing tasks, allowing them to focus on creating an enjoyable experience for the players. FIREBALL opens the door to multiple exciting avenues of research, and we're excited to see how future work utilizes our unique dataset of state-augmented gameplay.

8 Limitations

Dungeons & Dragons is a very complex game to capture completely, and there are certain aspects that FIREBALL does not take into account. For example, FIREBALL's scenarios are recorded independently of the overarching narrative context they take place in, do not record players' inventory, and do not account for any movement or placement on a map. Our models are not able to play D&D autonomously - but doing so is not the goal. Instead, D&D models can be used to assist and inspire the humans playing.

Our models do not take into account the gener-

ation of profanity or sensitive topics; these were filtered out post-hoc. D&D is a game played by players of all ages that often contains violent or profane descriptions, and unfiltered generations may be unsuitable for young players. There are previous instances of roleplaying games that incorporate language models being used to generate sexual content[3] that would require age restrictions and content warnings.

GPT-3 may be prohibitively expensive for everyday use; in our experiments, we were unable to use the full set of data we had available for fine-tuning due to budget constraints.

Acknowledgements

References

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and Dragons as a Dialogue Challenge for Artificial Intelligence. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding methods for neural narrative generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185, Online. Association for Computational Linguistics.

Simon Ellis and James Hendler. 2017. Computers Play Chess, Computers Play Go...Humans Play Dungeons & Dragons. *IEEE Intelligent Systems*, 32(4):31–34.

João Gabriel Gama Vila Nova, Marcos Vinícius Carneiro Vital, and Roberta Vilhena Vieira Lopes. 2019. A Reliable Information Acquisition Model for D&D Players. In *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, page 73–85, Cham. Springer International Publishing.

Gary Gygax and Dave Arneson. 1974. *Dungeons & Dragons*.

---

[3] https://www.theregister.com/2021/10/08/ai_game_abuse/

Apoorv Kulshreshtha, Daniel De Freitas Adiwardana, David Richard So, Gaurav Nemade, Jamie Hall, Noah Fiedel, Quoc V. Le, Romal Thoppilan, Thang Luong, Yifeng Lu, and Zi Yang. 2020. Towards a human-like open-domain chatbot. In *arXiv*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D. Ernst. 2018. NL2Bash: A corpus and semantic parser for natural language interface to the linux operating system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nicholas Locascio, Karthik Narasimhan, Eduardo DeLeon, Nate Kushman, and Regina Barzilay. 2016. Neural generation of regular expressions from natural language with minimal domain knowledge. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1918–1923, Austin, Texas. Association for Computational Linguistics.

Annie Louis and Charles Sutton. 2018. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *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. Association for Computational Linguistics.

Lara J. Martin, Srijan Sood, and Mark O. Riedl. 2018. Dungeons and DQNs: Toward Reinforcement Learning Agents that Play Tabletop Roleplaying Games. In *Joint Workshop on Intelligent Narrative Technologies and Workshop on Intelligent Cinematography and Editing (INT-WICED)*, Edmonton, AB, Canada. http://ceur-ws.org.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic. Association for Computational Linguistics.

Pax Newman and Yudong Liu. 2022. Generating Descriptive and Rules-Adhering Spells for Dungeons & Dragons Fifth Edition. In *Games and Natural Language Processing Workshop at LREC*, page 54–60, Marseille, France. European Language Resources Association.

Stefan Papazov, Wesley Gill, Marta Garcia Ferreiro, Andrew Zhu, Lara J. Martin, and Chris Callison-Burch. 2022. Using language models to convert between natural language and game commands. In *The Third Wordplay: When Language Meets Games Workshop*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online. Association for Computational Linguistics.

Manasvi Sagarkar, John Wieting, Lifu Tu, and Kevin Gimpel. 2018. Quality signals in generated stories. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 192–202, New Orleans, Louisiana. Association for Computational Linguistics.

Wai Man Si, Prithviraj Ammanabrolu, and Mark Riedl. 2021. Telling stories through multi-user dialogue by modeling character relations. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 269–275, Singapore and Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2022. An ai dungeon master's guide: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons.

A Example of Collected Character Information

Here, Table 6, we provide a table of all of the state and character information that we collected from actual play. The table gives an example for the character Filgo Bitterfoot, a level 5 Dwarf Fighter. We also explain the function of each attribute within the rules of D&D, as well as a possible use case for generative applications.

Attribute	Example	Description	Use Case
Name	Filgo Bitterfoot	Character's name.	Both commands and narration usually reference actors by their name.
Class Level	Fighter 5	Classes inform what types of actions the character is good at, while level is the amount of "experience" the character has.	An actor's classes offer a general overview of their abilities.
Stats	STR: 15; DEX: 10; CON: 17; INT: 10; WIS: 14; CHA: 10; Proficiency: +3	Strength, dexterity, constitution, intelligence, wisdom, & charisma. Being proficient in a stat or skill means that a bonus is added to their die outcome.	Can be used to condition text generation; for example, a strong character's actions would likely be described differently from a dexterous character.
Skills	Acrobatics +0; Animal Handling +5 (proficient);...		
Saves	STR +5; DEX +0; CON +6; INT +0; WIS +2; CHA +0	Saves are bonuses for when a character is defending against an attack.	
Resistances	Resistant to fire, poison	Cuts the amount of damage done or removes it entirely if they are *immune*.	Useful for game reasoning: don't use fire abilities against an immune target.
Attacks	Frost Brand Scimitar (+5 to hit, 2d6+2 damage); Unarmed Strike (+5 to hit, 1d4+2 damage); ...	Weapons, bonuses for the ability to hit the target, and the amount of damage the target takes when hit. 2d6 indicates 2 6-sided dice.	These attributes provide a more detailed list of the actor's capabilities. For command prediction, the actor may only use an ability they actually have. For text generation, an ability list can help describe the actor's style (e.g. "an ice mage").
Spellbook	+5 Spell Bonus; DC 13; Spells: Fireball, Prestidigitation, ...	Bonus to see if a spell hits the target, the amount the target needs to beat when being hit by the spell in order to dodge it, a list of spells that the character can cast (each with different outcomes)	
Actions[†]	Rally; Second Wind; Interception; Action Surge; Disarming Attack; ...	Class-specific abilities.	
Custom Counters[†]	Superiority Dice: 4/4; Second Wind: 1/1; Action Surge: 1/1	Limitation on the number of times the character can use an action before resting.	Certain abilities can only be used a certain number of times.
Armor Class	18	How difficult the character is to hit.	Provides information about how dire a situation might be for more interesting text generation. Actors are also commonly referred to by a combination of these (e.g. "the prone dwarf").
Hit Points	54 / 54 (0 temp)	How "healthy" a character is. Filgo has 54 points out of 54.	
Effects[*]	Prone, Stunned	Combat-specific states that affect how attacks resolve.	
Creature Type	Humanoid	Category of creature. Other examples are *undead* and *dragon*.	
Race[†]	Mountain Dwarf	Fantasy race that might provide extra bonuses or actions.	
Description[†]	"Filgo is a level 5 Mountain Dwarf Fighter. He is 100 years old, 4' tall..."	Natural-language description of the character, how they look & what they're like.	Provides qualitative information about the actor useful for text conditioning, e.g. Louis and Sutton (2018).

Table 6: A list of common attributes associated with actors in FIREBALL. [†]Only available for player characters. [*]Only available for actors in combat. Character sheet for this example actor available at https://github.com/zhudotexe/FIREBALL/blob/main/Filgo_Bitterfoot.pdf.

Event	Count	Description
`message`	8,012,706	A user has sent a message in the recorded Discord channel.
`command`	2,109,603	An Avrae command has successfully executed. Includes information about the game actor that ran the command.
`combat_state_update`	1,297,254	Some element of the combat state has changed. Includes the new combat state.
`automation_run`	588,712	An action has successfully executed. Includes information about all the rolls in and result of the action, including successes and failures of attacks, saves, etc.
`alias_resolution`	325,860	A user's custom command finished executing. Includes the command's code, the message content before it was run, and the updated command.
`snippet_resolution`	102,730	A user's custom argument shortcut finished executing. Includes the shortcut's code, the argument list before it was run, and the updated argument list.
`combat_start`	24,748	A user started combat in a channel.
`combat_end`	23,469	A user ended combat in a channel. (Combats left inactive for over a month are automatically ended without emitting an event.)
`button_press`	21,756	A user pressed a button associated with a game effect, such as the "Stand Up" button associated with the Prone effect.

Table 7: The number of events in the dataset.

B Dataset Events

Table 7 lists all the events in the dataset along with their descriptions. See `https://github.com/avrae/avrae/blob/v4.2.2/cogs5e/initiative/upenn_nlp.py` for each event's schema.

C Action Attributes

Actions consist of a tree of *effects*, such as rolling to hit, dealing damage, or rolling a saving throw. Table 8 lists many relevant effects and the attributes available in each. Further documentation about actions is available at `https://avrae.readthedocs.io/en/latest/automation_ref.html` and the full definition is available at `https://github.com/avrae/avrae/blob/v4.2.2/cogs5e/models/automation/results.py`.

Effect	Attribute	Description
attack	did_hit	*Whether or not an attack roll hit its target.*
	did_crit	*Whether or not an attack roll was a critical hit.*
save	dc	*The difficulty class of this saving throw.*
	ability	*The ability this saving throw uses.*
	did_save	*Whether the saving roll was successful.*
damage	damage	*The total amount of damage dealt to a target.*
	in_crit	*Whether the damage dealt was part of a critical hit.*
temphp	amount	*The total amount of temporary hit points granted to a target.*
ieffect	effect	*A temporary game effect was placed on a target (e.g. Prone).*
remove_ieffect	effect	*A temporary game effect was removed from a target.*
check	skill_name	*The name of a skill the target rolled for a check.*
	dc	*The difficulty class of this ability check, if applicable.*
	did_succeed	*Whether the check was against a DC and the target succeeded.*
	contest_roll	*The result of the caster's contesting ability check, if applicable.*
	contest_did_tie	*Whether the both actors rolled the same result in a contest.*

Table 8: The attributes associated with each effect in an action.

D Utterance To Command Generation Perturbation Details

For the perturbation experiments, we selected 2 specific scenarios, one in which the player chooses to cast a bardic inspiration spell to target a single member of their party and the other where the player chooses to cast Fireball, perhaps the most canonical spell in D&D. For both of these scenarios; we took the combat

state and prompt from their respective unit tests and then pertubed them in order to test the ability of the model to react to various modifications in input. We seek to study the model responses specifically to the following scenarios

- Targeting and association - can the model pick up intended targets from nicknames/character classes/races in the input. Eg, Can it determine that "Inspires the druid" and "Inspires Noxxis" should result in targeting the same character (if Noxxis is a druid)?
- Can it recognize a spell from a creative description of its effects?
- Does it attend to spells in the prepared spell list?
  While we do not perform exhaustive quantitative analysis, a preliminary analysis indicates that the Finetuned model that includes the state information can react to changes in the prepared spell list.

```
"Actors:
    - OR2 (Orc) <9/15 HP; Injured>
    - Reef (Variant Human; Sorcerer 1/Bard 2) <19/25 HP; Injured>
    - KO3 (Kobold) <5/5 HP; Healthy>
    - Rahotur (Mountain Dwarf; Barbarian 6) <77/77 HP; Healthy>
    - Calti Xihooda (Lizardfolk; Druid 6) <45/45 HP; Healthy>
    - Noxxis Blazehammer (Hill Dwarf; Cleric 7) <59/59 HP; Healthy>
    - OR1 (Orc) <13/15 HP; Injured>
    - KO1 (Kobold) <3/5 HP; Injured>
    - KO2 (Kobold) <5/5 HP; Healthy>
    - OR3 (Orc) <15/15 HP; Healthy>
    - OR4 (Orc) <15/15 HP; Healthy>
Current:
    Name: Noxxis Blazehammer
    Class: Cleric 7
    Race: Hill Dwarf
    Attacks: Warhammer, 2-Handed Warhammer, Unarmed Strike
    Spells: Fireball, Death Ward, Word of Radiance, Hold Person, Spiritual
        Weapon, Revivify, Augury, Scorching Ray, Light, Healing Word, Spirit
         Guardians, Guidance, Burning Hands, Faerie Fire, Guiding Bolt,
        Flaming Sphere, Thaumaturgy, Cure Wounds, Bless, Protection from
        Evil and Good, Daylight, Wall of Fire, Sacred Flame, Guardian of
        Faith
    Actions: Channel Divinity, Warding Flare, Channel Divinity: Radiance of
        the Dawn, War Caster, Channel Divinity: Turn Undead, Destroy Undead
        , Harness Divine Power

    *Noxxis invokes divine anger of his deity, coalescing it into a gout of
        flame that he launches towards the orcs*",
```

This prompt generates the command *!cast fireball -t OR1 -t OR2 -t OR3 -t OR4* but if the Fireball spell is removed from the prompt, it generates the command *!cast "burning hands" -t or1 -t or2 -t or3 -t or4* Similarly, in the case of the Bardic Inspiration example, it's able to replace Bardic Inspiration with Healing Word. The model seems to be able to reliably differentiate between healthy and injured enemies - asking the model to cast Fireball at the injured enemies generates the appropriate command. It also seems to be able to target based on character classes and races. However, it does not always target the correct number of enemies - asking the model to target "2 injured orcs" leads to targeting all the injured orcs. Similarly, asking the model to target based on party roles fails; asking the model to target "the casters" does not correctly target the spellcasters in the party. While these results are promising, we leave exhaustive quantitative evaluation to future work.

E Full State to Narration Prompt

```
History:
Player 3: (thunder, if it matters?)
Player 3: As the blast hits the hag, another Wild Magic Surge bursts from the gobbo
Player 3: And they turn blue. They look like a Verdan now, and they feel slight
    amounts of shame from it
Player 3: But, the day goes on, and the reach out with a spectral hand to backhand
    slap the hag with... Chill Touch, imbuing it with the Tides of Chaos for a
    little extra sting
Player 4: Well, looks like he got up fine by himself...

There's more bashing to do. Raising the head of the mace, she brings it down once
    again.
---

Actors:
- Verity Silverdust (Halfling; Rogue 3) <18/18 HP; Healthy> [Mage Armor]
- Nitar (Variant Human; Barbarian 3) <1/35 HP; Critical> [Frightened, Wildhunt
    Shifting, Rage]
- Bartholomew (Goblin; Sorcerer 3) <23/23 HP; Healthy> [Wild Resistance, Chilling
    Touch]
- Alexsandra (Astral Elf; Cleric 3) <15/15 HP; Healthy>
- Keya (Custom Lineage; Fighter 2/Warlock 1) <24/24 HP; Healthy> [Hexblade's Curse,
     Hex, Hexing]
- Mozzie Urahaka (Dhampir; Artificer 1/Wizard 2) <22/22 HP; Healthy> [Mind Splinter
    ]
- SH1 (Sea Hag) <2/52 HP; Critical> [Hexblade's Cursed, Chill Touch, Hexed]

Targets:
- SH1 (Sea Hag) <2/52 HP; Critical> [Hexblade's Cursed, Chill Touch, Hexed]

Description:
A timeless woman with flowing star-speckled hair that fades between nebula violet
    and empty black, she wears and carries exotic armor and weaponry: most of all
    being the amulet of a horned clawed cyclopic serpent swallowing its own tail.
    She manages the altar and gravestones at the cemetery and ensures that no
    desecration comes to those under her care. When spoken to, the gravekeeper has
    an odd aura about her, being generally much too wide-eyed and profound for
    simple small-talk.
---
Name: Alexsandra
Class: Cleric 3
Race: Astral Elf
Attacks: Crossbow, light, Mace, Unarmed Strike
Actions: Channel Divinity: Radiance of the Dawn, Channel Divinity: Turn Undead,
    Harness Divine Power, Starlight Step, Warding Flare
Spells: Dancing Lights, Guidance, Light, Sacred Flame, Spare the Dying, Thaumaturgy
    , Bless, Burning Hands, Command, Faerie Fire, Healing Word, Sanctuary,
    Blindness/Deafness, Flaming Sphere, Gentle Repose, Lesser Restoration,
    Scorching Ray, Silence
```

```
Aleksandra attacks with a Mace!
Aleksandra attacked SH1 but missed.
```

F Full Utterance to Command Prompt

```
Actors:
- Filgo Bitterfoot (Mountain Dwarf; Fighter 5) <43/43 HP; Healthy>
- DW1 (Dire Wolf) <25/37 HP; Injured>

Current:
Name: Filgo Bitterfoot
Class: Fighter 5
Race: Mountain Dwarf
Attacks: Greataxe, Longsword, Longbow, Handaxe
Actions: Second Wind, Action Surge

Filgo swings his axe at the wolf! "Raaaargh!"
```

For this example, the true command is:

```
!a greataxe -t dw1
```

G Human Evaluation Interface

G.1 Signup Form

To assess each rater's familiarity with Avrae and D&D, we asked each the following questions:

1. Have you ever played the fifth edition of Dungeons and Dragons (D&D 5e) using the Avrae Discord Bot before?

2. If so, roughly how long have you been playing using Avrae?

3. Have you ever been the Dungeon Master of a D&D 5e game using Avrae?

4. If so, roughly how long have you been DMing using Avrae?

5. Have you used any of these commands on the Avrae Discord Bot?

   (a) !import, !check, !save, !action, !cast
   (b) !init begin, !init next, !init join, !init action

6. Have you ever played on a "play-by-post" D&D Discord?

7. What's one feature you always wished was in the Avrae Discord Bot?

G.2 Evaluation Mockup

In this task, you will see part of a play-by-post D&D combat using Avrae in the form of Discord messages leading up to an Avrae action. The caster's description and current initiative list are listed along with the Discord messages. The messages that are shown as context are real messages from players. Your job is to read the context and then rate different responses for the dungeon master's narration of the action. Please note that the context you are given represents only a part of the players' past conversations/interactions with one another during the game.

Caster's Description

Description: 5'10 (180cm) | 180 lb. | Chromatic Dragonborn | Fighter (Battle Master)/Bard

Young, lean but strong overall build. They're a blue chromatic dragonborn who's always seen in armor and formal decorated robes. With silvery blond hair that is usually hidden behind a helm. Umbrage has many choices of weaponry, not one to pick or choose when it comes to the field of battle. But his most favored would be that horn of his. A rustic and old warhorn, The ivory it's made from is something unusual, even going so far as to be able to tap into the wave around him by chance.

With the passing battles, many new scars are shown upon scales. But the only one that bothers him the most and that is always is kept hidden. Is the injury found upon his neck. The cause must have been something heavy enough to leave a lasting imprint, but Umbrage would never tell what it was. Shocking anyone who gets too close to that reverse scale.

Initiative List

- DD6 (Death Dog) <11/39 HP; Bloodied> [baki]
- Katherine (Dhampir; Rogue 2/Blood Hunter 3) <46/46 HP; Healthy>
- Holawynn Meitorin (Satyr; Cleric 5) <48/48 HP; Healthy>
- DD1 (Death Dog) <39/39 HP; Healthy> [Umbrage]
- DD2 (Death Dog) <-19/39 HP; Dead> [yala]
- Kaska (Leonin; Ranger 7/Paladin 3) <114/114 HP; Healthy>
- Baki (Beast of the Land) <33/40 HP; Injured> [Maul, Primal Bond, Poisoned]
- DD7 (Death Dog) <39/39 HP; Healthy> [Lytrha]
- Umbrage (Chromatic Dragonborn; Fighter 3/Bard 4) <56/63 HP; Injured> [Cause Fear]
- Lythra (Half-Orc; Ranger 3/Rogue 2) <42/42 HP; Healthy>
- Yala the Wanderer (Goblin; Bard 5/Rogue 1) <39/39 HP; Healthy>
- DD8 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD3 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD5 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD4 (Death Dog) <39/39 HP; Healthy> [40 feet]

Context

Avrae | 3 minutes ago
Initiative 12 (round 1): Umbrage ()

Umbrage <56/63 HP> (AC 20)
> Resistances: lightning
> Immunities: critical hits, diseas

Player 0
 a minute ago
"Dogs... so many of them making it more annoying then anything." Umbrage didn't want to waste anything big. So he pulled out his pipes, taking the attack of opportunity as he would start to play.

"Fear me!"

Player 0 | now
!i cast "Cause Fear" -dc 15 -t dd1 -t dd3 -t dd4 -t dd5 -t dd6 -t dd7 -t dd8 sadv -i -title "Pipes of Haunting!" -thumb https://cdn.discordapp.com/attachments/979075315706003459/1005867411385364480/giphy_6.gif

```
Avrae | now
Pipes of Haunting!
Meta
DC: 15

DD1
WIS Save: 2d20kh1 (15, 12) + 1 = 16; Success!

DD3
WIS Save: 2d20kh1 (5, 2) + 1 = 6; Failure!
Effect: Frightened (Cause Fear) [10 rounds] (Resist Fear; Parent: Cause Fear)

DD4
WIS Save: 2d20kh1 (9, 16) + 1 = 17; Success!

DD5
WIS Save: 2d20kh1 (2, 8) + 1 = 9; Failure!
Effect: Frightened (Cause Fear) [10 rounds] (Resist Fear; Parent: Cause Fear)

DD6
WIS Save: 2d20kh1 (19, 18) + 1 = 20; Success!

DD7
WIS Save: 2d20kh1 (20, 17) + 1 = 21; Success!

DD8
WIS Save: 2d20kh1 (9, 3) + 1 = 10; Failure!
Effect: Frightened (Cause Fear) [10 rounds] (Resist Fear; Parent: Cause Fear)

Effect
You awaken the sense of mortality in one creature you can see within range. A
    construct or an undead is immune to this effect. The target must succeed on a
    Wisdom saving throw or become frightened of you until the spell ends. The
    frightened target can repeat the saving throw at the end of each of its turns,
    ending the effect on itself on a success.
```

Does the response make sense?

Use your common sense here. The response "makes sense" if:

1. it is cohesive as a standalone statement,

2. consistent with the rules of the game,

3. and it is a plausible narration given the prior context (initiative list and last actions taken).

If anything seems off—not fluent, confusing, illogical, out of context, or wrong according to the rules of D&D —then choose No. If in doubt about a response, choose No.

	Yes	No
The music was certainly not pleasant, but it seemed to have an effect on the dogs. They became panicked for a moment, but were able to maintain their aggression. It was just enough for the party to act.	☐	☐
Cuco proceeds to try and scare the ducks. So far it wasn't working.	☐	☐
Moving back into the melee with the one he was originally fighting. As he drew ready.	☐	☐
As the first dog fell, the others seemed to flee once more in pure terror of the seemingly unending tunes of the pipes.	☐	☐
The dogs were in the middle of a panic. But still they fought on!	☐	☐

Is the response specific?

In other words, do you think that the response accurately narrates the last action the character actually took and its results?

The response is "specific" if it flows logically from the specific action and result taken by the character, in the greater context provided.

Note: It is possible for a response to "make sense" (due to being cohesive, consistent and plausible in and of itself), but be marked "not specific" when it is not a logical next step in the overall game progression.

Note: "Specific" for the purposes of this task does not have to do with how detailed the response is per se; a response can be fairly general in its language, but still qualify as "specific" when it is a logical next step in the overall game progression.

	Yes	No
As the first dog fell, the others seemed to flee once more in pure terror of the seemingly unending tunes of the pipes.	☐	☐
Cuco proceeds to try and scare the ducks. So far it wasn't working.	☐	☐
Moving back into the melee with the one he was originally fighting. As he drew ready.	☐	☐
The dogs were in the middle of a panic. But still they fought on!	☐	☐
The music was certainly not pleasant, but it seemed to have an effect on the dogs. They became panicked for a moment, but were able to maintain their aggression. It was just enough for the party to act.	☐	☐

How interesting is the response? (10 is best)

Rank a response as more "Interesting" if the response would likely catch someone's attention or arouse curiosity in the game; or it is insightful, creative, or witty with respect to the game. If the response is monotonous and predictable, then rank it lower. If anything seems off—not fluent, confusing, illogical, out of context, or wrong according to the rules of D&D —then rank it lower.

Less Interesting									More Interesting
1	2	3	4	5	6	7	8	9	10

The music was certainly not pleasant, but it seemed to have an effect on the dogs. They became panicked for a moment, but were able to maintain their aggression. It was just enough for the party to act.

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

The dogs were in the middle of a panic. But still they fought on!

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Cuco proceeds to try and scare the ducks. So far it wasn't working.

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

As the first dog fell, the others seemed to flee once more in pure terror of the seemingly unending tunes of the pipes.

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Moving back into the melee with the one he was originally fighting. As he drew ready.

○ ○ ○ ○ ○ ○ ○ ○ ○ ○

H Human Evaluation Significance

We use the Student's t-test to calculate significance for our three rated dimensions. The results are tabulated below, with **bold** indicating $p < 0.001$, *italics* indicating $p < 0.01$, and † indicating $p < 0.05$:

Significance of Sense (p-value)

	FIREBALL-FULL	FIREBALL-SHORT	COMMAND	DIALOG
FIREBALL-SHORT	0.2896	-	-	-
COMMAND	**0.0002**	*0.0046*	-	-
DIALOG	**0.0000**	**0.0000**	0.1410	-
Human	0.5995	0.6179	*0.0014*	**0.0000**

Significance of Specific (p-value)

	FIREBALL-FULL	FIREBALL-SHORT	COMMAND	DIALOG
FIREBALL-SHORT	0.7412	-	-	-
COMMAND	*0.0079*	*0.0025*	-	-
DIALOG	**0.0000**	**0.0000**	*0.0060*	-
Human	0.8670	0.8651	*0.0059*	**0.0000**

Significance of Interesting (p-value)

	FIREBALL-FULL	FIREBALL-SHORT	COMMAND	DIALOG
FIREBALL-SHORT	0.0602	-	-	-
COMMAND	0.5189	0.1996	-	-
DIALOG	0.0710	**0.0001**	$0.0227^{\dagger}$	-
Human	0.1288	0.7277	0.4153	*0.0019*

I Example of Full State to Narration Context

The following is the full prompt provided to the FIREBALL-FULL model in the example provided in Table 5.

```
History:
Player 1: *Holawynn would back up 35 if she could. The satyr knew where she should
    be standing during this fight. She starts up with a twilight flame to conserve
    slots.*
Player 1: *It misses out of sheer unluck. A shame that it was also kinda crap.*
```

Player 0: ```The hounds charged. Each managing a singular bite on their targets. It
    would seem they were all wanting to eat some tender flesh of the adventures
    who passed through their masters lair!```
Player 2: *Kaska looked about the combat and swung her weapon to Yala's aid, Baki
    attacking the dog that wanted to eat his bacon.*
Player 0: "Dogs... so many of them making it more annoying then anything." *Umbrage
    didn't want to waste anything big. So he pulled out his pipes, taking the
    attack of opportunity as he would start to play.*
"**Fear me!**"

Actors:
- DD6 (Death Dog) <11/39 HP; Bloodied> [baki]
- Katherine (Dhampir; Rogue 2/Blood Hunter 3) <46/46 HP; Healthy>
- Holawynn Meitorin (Satyr; Cleric 5) <48/48 HP; Healthy>
- DD1 (Death Dog) <39/39 HP; Healthy> [Umbrage]
- DD2 (Death Dog) <-19/39 HP; Dead> [yala]
- Kaska (Leonin; Ranger 7/Paladin 3) <114/114 HP; Healthy>
- Baki (Beast of the Land) <33/40 HP; Injured> [Maul, Primal Bond, Poisoned]
- DD7 (Death Dog) <39/39 HP; Healthy> [Lytrha]
- Umbrage (Chromatic Dragonborn; Fighter 3/Bard 4) <56/63 HP; Injured> [Cause Fear]
- Lythra (Half-Orc; Ranger 3/Rogue 2) <42/42 HP; Healthy>
- Yala the Wanderer (Goblin; Bard 5/Rogue 1) <39/39 HP; Healthy>
- DD8 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD3 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD5 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD4 (Death Dog) <39/39 HP; Healthy> [40 feet]

Targets:
- DD1 (Death Dog) <39/39 HP; Healthy> [Umbrage]
- DD3 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD4 (Death Dog) <39/39 HP; Healthy> [40 feet]
- DD5 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]
- DD6 (Death Dog) <11/39 HP; Bloodied> [baki]
- DD7 (Death Dog) <39/39 HP; Healthy> [Lytrha]
- DD8 (Death Dog) <39/39 HP; Healthy> [40 feet, Frightened (Cause Fear)]

Description: __**5'10 (180cm) | 180 lb. | Chromatic Dragonborn | Fighter (Battle
    Master)/Bard**__
> Young, lean but strong overall build. They're a blue chromatic dragonborn who's
    always seen in armor and formal decorated robes. With silvery blond hair that
    is usually hidden behind a helm. Umbrage has many choices of weaponry, not one
    to pick or choose when it comes to the field of battle. But his most favored
    would be that horn of his. A rustic and old warhorn, The ivory it's made from
    is something unusual, even going so far as to be able to tap into the wave
    around him by chance.
>
> With the passing battles, many new scars are shown upon scales. But the only one
    that bothers him the most and that is always is kept hidden. Is the injury
    found upon his neck. The cause must have been something heavy enough to leave a
    lasting imprint, but Umbrage would never tell what it was. Shocking anyone who

gets too close to that reverse scale.
Personality Traits
I'm haunted by memories of war. I can't get the images of violence out of my mind.
I'm full of inspiring and cautionary tales from my military experience relevant to
    almost every combat situation.
I can stare down a hell hound without flinching.
Ideals
Might. In life as in war, the stronger force wins. (Evil)
Bonds
My honor is my life.
I'll never forget the crushing defeat my company suffered or the enemies who dealt
    it.
Flaws
I made a terrible mistake in battle that cost many lives-and I would do anything to
    keep that mistake secret.
Alignment
NE

Name: Umbrage
Class: Fighter 3/Bard 4
Race: Chromatic Dragonborn
Attacks: Javelin of Lightning, Light Hammer, +1, Longsword, +1, 2-Handed Longsword,
    +1, Warhammer, +1, Unarmed Strike, Keoghtom's Ointment
Spells: Ice Storm, Hold Person, Sleet Storm, Vicious Mockery, Faerie Fire, Ray of
    Frost, Thunderclap, Healing Word, Bane, Dissonant Whispers, Silence, Sleep,
    Prestidigitation
Actions: Maneuvers: Commander's Strike, Superiority Dice, Action Surge, Maneuvers:
    Bait and Switch (Self), Maneuvers: Bait and Switch (Target), Combat Inspiration
    , Song of Rest, Martial Adept, Maneuvers: Distracting Strike, Second Wind,
    Bardic Inspiration, Lightning Breath Weapon, Maneuvers: Ambush
Effects: Cause Fear

Pipes of Haunting!
DD1 rolled a Wisdom save and succeeded.
DD3 rolled a Wisdom save but failed.
DD3 gained Frightened (Cause Fear).
DD4 rolled a Wisdom save and succeeded.
DD5 rolled a Wisdom save but failed.
DD5 gained Frightened (Cause Fear).
DD6 rolled a Wisdom save and succeeded.
DD7 rolled a Wisdom save and succeeded.
DD8 rolled a Wisdom save but failed.
DD8 gained Frightened (Cause Fear).