Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence

Chris Callison-Burch*
University of Pennsylvania
ccb@upenn.edu

Gaurav Singh Tomar*
Google
gtomar@google.com

Lara J. Martin
University of Pennsylvania

Daphne Ippolito
University of Pennsylvania
Google

Suma Bailis
Google

David Reitter
Google

Abstract

AI researchers have posited Dungeons and Dragons (D&D) as a challenge problem to test systems on various language-related capabilities. In this paper, we frame D&D specifically as a dialogue system challenge, where the tasks are to both generate the next conversational turn in the game and predict the state of the game given the dialogue history. We create a gameplay dataset consisting of nearly 900 games, with a total of 7,000 players, 800,000 dialogue turns, 500,000 dice rolls, and 58 million words. We automatically annotate the data with partial state information about the game play. We train a large language model (LM) to generate the next game turn, conditioning it on different information. The LM can respond as a particular character or as the player who runs the game—i.e., the Dungeon Master (DM). It is trained to produce dialogue that is either in-character (roleplaying in the fictional world) or out-of-character (discussing rules or strategy). We perform a human evaluation to determine what factors make the generated output plausible and interesting. We further perform an automatic evaluation to determine how well the model can predict the game state given the history and examine how well tracking the game state improves its ability to produce plausible conversational output.

1 Introduction

Artificial Intelligence has a long and rich history of using games as challenge problems that lead to advances in the field. In many cases, AI game-playing systems have gone on to rival human champions of the game.

Tackling board games like checkers, backgammon, and chess resulted in AI search algorithms like MiniMax and alpha-beta pruning as well as representations of the search space like game trees, and heuristic evaluation on non-terminal game state.

*Denotes equal contribution

1992	IBM's TD-Gammon becomes Backgammon champ using temporal-difference learning
1994	University of Alberta's Chinook checker player declared world champion
1997	IBM's Deep Blue beats Garry Kasparov chess grandmaster
2011	IBM Watson Beats Ken Jennings at Jeopardy
2013	DeepMind "Playing Atari with Deep Reinforcement Learning"
2016	DeepMind's AlphaGo beats Lee Sedol in a five-game match
2019	DeepMind AlphaStar becomes a grandmaster StarCraft player
2019	OpenAI Five defeats world champion DOTA 2 players

Table 1: AI game playing systems have surpassed human champions for many games

Video games like StarCraft or classic Atari games have provided test beds for reinforcement learning where systems must learn game playing policies by interacting with dynamic worlds where the game state is not as easily repressed as in board games (Vinyals et al., 2019; Mnih et al., 2013). Trivia games like Jeopardy or Quiz Bowl have presented language-related challenges that advanced question-answering and information extraction (Ferrucci et al., 2010; Iyyer et al., 2014).

1.1 Challenge Problem for AI: Role Playing Games

Dungeons and Dragons has been identified as an appropriate challenge for the next stage of artificial intelligence (Ellis and Hendler, 2017; Louis and Sutton, 2018; Martin et al., 2018b). Ellis and Hendler (2017) proposed open-ended creative games like D&D as the next challenge for AI after the human-level successes by AI at Chess and Go, which are zero-sum, deterministic, sequential two-player games with perfect information. Louis and Sutton (2018) understood the importance of narrative in NLP/NLG and saw how cooperative story generation between humans already exists in these games and can be used for automated generation. Martin et al. (2018b) took this a step further and outlined some of the specific challenges D&D presents to the NLP community; such as a state of the game world distributed across the Dungeon Master (DM) and other players or dealing with the intrinsic rewards players get from taking certain actions that would not necessarily provide them with points or experience within the game.

Role playing games like Dungeons and Dragons are an interesting challenge problem for AI. Gameplay happens through language rather than moves on a game board. D&D involves multiple players who roleplay characters in a fantasy setting, guided by a Dungeon Master who sets obstacles and adventures and plays monsters. To have an AI successfully play D&D requires abilities like

- Language generation (multi-party dialog, generating descriptions of the world/actions, storytelling)
- Language understanding (knowledge acquisition and representation, state tracking, automated reasoning)
- Planning / strategic play during battles (similar to chess or go)

Is it possible to design an AI system that is capable of playing a game of D&D either as a character in the game or as the Dungeon Master using current AI technology?

1.2 Why is this the right time for this challenge?

Large scale neural language models like GPT have shown impressive generation results (Brown et al., 2020). Incorporating Neural LMs into a game setting exercises their strengths and exposes their weaknesses. D&D -style role playing games are a mix of language generation, language understanding, state tracking, and rule following that make them a good research challenge that could advance AI research.

1.3 Research Challenges Presented by Role Playing Games

Instead of the game being a series of moves on a game board, Role Playing Games (RPGs) like Dungeons and Dragons (D&D) are language-based. Players create characters that have a class (wizard, fighter, thief, etc.) that denotes their abilities, and a fantasy race (elf, dwarf, human, etc.). Players describe what they want their character to do and roll dice to determine if they are successful. The dungeon master (DM) acts as the narrator who shapes the overall story. The DM describes scenarios and locations, and takes on the role of non-player characters (NPCs), and monsters.

A common element to the game play is an encounter with monsters. Battles are governed by rules, and unfold in a turn-based fashion where the DM controls the monsters and each player controls their character. Each player and monster has a health meter (called their HIT points), an armor class (which indicates the threshold of the dice roll needed to damage them), and a set of possible attack or move actions.

Table 2 provides example dialogue from a game of D&D being played between 3 players – Travis (playing a human fighter named Magnus Burnsides), Clint (playing Merle Highchurch, a dwarf cleric), Justin (playing Taako an elf wizard), and DM Griffin. We add comments about each dialogue turn to describe what is happening in the game, and to highlight the challenges that would need to be addressed if an AI system were to play the game either as a player or as the DM.

1.4 Our Contributions

In this paper, we introduce a new dataset of "actual play" game transcripts. Each turn is labeled with state variables, character information, partial game state, and whether the conversational turn was in-character or out-of-character. Our data is a novel, large scale, real-world conversational dataset. It is unique in that the dialog turns are generated entirely through player collaboration and written interaction. Unlike existing dialog datasets, our data is modeling the Dungeons & Dragons role-playing game as a multi-party dialogue. We also train a large language model to perform response generation and game state tracking. Our dataset is interesting as a challenge for dialogue systems for the following reasons:

- It is naturally occurring dialog that is purely conversational.
- It is strongly history dependent – a substantive criticism of recent dialog datasets is their history independence (Mosig et al., 2020).
- It covers a spectrum of task oriented and non-task oriented (e.g. chit chat) dialog.

Player (character)	**Game Dialogue**	**D&D Game Description and AI challenges**
Griffin (DM)	A dwarf named Gundren Rockseeker has hired you to transport a wagonload of provisions to the rough-and-tumble settlement of Phandalin, which is a couple days' travel to the southeast. A day and a half after leaving, you turn off the high road that connects the major cities on the coast onto a smaller trail that will lead you to Phandalin. This trail is not as well maintained, and bandits and outlaws have been known to lurk along the trail.	*This game is based on the D&D starter adventure called "Lost Mine of Phadelver". The adventure book is a mixture of rules and "boxed text" which is descriptive text for the DM to read aloud or paraphrase. See the appendix for the text that the DM is consulting.* AI challenges: **Generation of stories and descriptive text**
Griffin (DM)	Roll a perception check for me. Perception is a wisdom skill, so be sure to add your wisdom modifier.	*The previous text was descriptive text. Here the DM is asking the players to perform a game mechanic and referencing a game rule. This is called "out of character" dialogue.* AI challenges: **Knowledge base population (extraction of rules from a rulebook)**
Clint (out of character)	I got an eight.	*Clint has rolled his dice. The number is low so his character fails the check.* AI challenges: **Multi-party dialogue**
Justin (out of character)	I got a six.	*Justin also fails. Neither character sees the thing that requires the perception check.*
Travis (out of character)	I rolled a natural twenty plus my wisdom modifier is 23.	*Travis rolls high number and succeeds on the check.* AI challenges: **Understanding rules, determining success or failure**
Griffin (DM)	With his eagle eyes, Magnus spots two dead horses lying in the middle of the road about 200 feet ahead of you.	*The DM describes what happens as a result of the success. AI challenges:* **Reasoning about consequences of success or failure, descriptive text generation**
Travis (in-character as Magnus)	I stop the wagon and motion silently to get the attention of Merle and Taako, and kinda pull them up towards the front of the wagon.	*Travis is describing what he is doing using "in character" language.* AI challenges: **Persona-based chat**
Griffin (DM)	As you warn them that shit has gone south, you notice a few goblins crouching in a part of the shaded woods off to the side of the road. Two of the goblins begin charging your wagon.	*The DM describes the start of a battle with several monsters.* AI challenges: **State tracking (in combat v. out of combat).**
Travis (out of character)	How many goblins are there?	AI challenges: **Question answering, state tracking (how many monsters).**
Griffin (DM)	There are three goblins; two of them are rushing the group, one is pretty heavily obscured by the brush, probably about 40 feet out, sort of between you and the dead horses laying in the middle of the road.	AI challenges: **Question answering, Descriptive text generation from game state.**
Clint (Merle)	I will cast sacred flame at the nearest one. If it fails a dexterity saving throw, it takes 6 points of damage.	*Clint chooses an action based on what is allowed for his character class. He describes the rule that governs the spell in an out-of-character fashion.* AI challenges: **Intent detection (perform attack action against a particular goblin)**
Griffin (DM)	You attack. You launch some fire onto the goblin closest to the wagon. And with that, he looks like he is on death's door. And the other goblin that you can see, the one that's not in the brush somewhere, just sort of stops in his tracks. What do you do next?	*The DM rolls for the monster, updates the state of its health meter, and describes the result of Merle's attack.* AI challenges: **Reasoning about rules, state tracking (monster's HIT points), descriptive text generation.**

Table 2: Example dialogue from a game of D&D with explanations of what is happening and comments on potential challenges for AI

- It has many participants in the conversation, since there are several players in the game.
- It conveys narrative elements including descriptions of events, that denote changes in the state of the game.

We evaluate a state-of-the-art language model, to understand how conditioning generation on different elements improves the quality of the generated text.

2 Tasks

We trained a large language model (LLM) to perform two tasks: **Next Utterance Prediction** and **Game State Tracking**.

Next Utterance Prediction We trained our language model on a corpus of human conversations (see Datasets section) to predict the next utterance. We varied the conditioning information to examine the effects on the quality of predicted next utterance. In all variations, we included the conversational history as input. We used most recent 7 conversational turns as input (in our example in the Table above, that would be from Griffin saying "With his eagle eyes…" up until "What do you do next?"). Given the conversational input (and other input in the variant models), the LLM must generate the next utterance, such that it is both interesting and a plausible next turn in the D&D game.

Game State Tracking In this task, rather than outing the next utterance, we had the model predict the game state for a given dialogue turn in the conversation. We have kept the state definition similar to task-oriented dialogue state tracking (DST). In DST, the dialogue state is a collection of slot-value pairs. In our case, each slot is a state variable feature related to D&D games. Our target slot values do not need to appear as a word in the dialogue context. We track several game states aspects including some that remain relatively static throughout the game (character attributes like their pronouns, class, fantasy race, and their inventory), some that change periodically (like being in combat or out of combat), and some that change from turn to turn (like what action the player is taking).

In this paper we attempt to track the following game state variables:

- Character specific state variables (remain constant throughout the game). These include Player ID, Character's name, pronouns, class and fantasy race, Items in character's inventory.
- Combat related state variables (change periodically) These include Whether the players are engaged in combat, Monsters in combat
- Dialog turn level state variables. These are actions related to dice rolls .

3 Datasets

For this paper, we have created a novel dataset for our dialogue-oriented test of AI's ability to play Dungeons & Dragons. We investigated two sources of data of people playing the game. The first source of data was "Actual Play" podcasts, where people record themselves actually playing RPGs. The second Play-By-Post data scraped from a web forum where people play by taking turns posting on the forum to describe their move. Ultimately, we focused primarily on the Play-By-Post data rather than the Actual Play Podcast data, since our automatic transcripts of the podcast audio had some speaker diarization errors and misrecognized words. Furthermore, using data from podcasts is not straightforward due to the inclusion of extraneous content, such as advertisements (Reddy et al., 2021).

3.1 Actual Play Podcasts

Actual Play podcasts are a genre of podcasts where people record themselves playing RPGs. We collected lists of popular actual play podcasts from the web, and expanded the list by searching for additional shows with relevant keywords including Actual Play, Role Playing Game, D&D 5e. Extracted all podcast episodes for these candidate podcast shows, and transcribed all podcast episodes using Google Cloud Speech-to-Text API (discarding about 15% of the episodes because of download errors or transcription failures). Table 3 shows a fragment of one of the podcast transcript. Table 4 summarizes the amount of Actual Play podcast data that we collected.

Dealing with speech data presents several challenges, including large audio files (the mean episode length was 1 hour and 17 minutes, with a max of 7 hours and 23 minutes), automatic speech recognition errors, and speaker diarization errors (2.2 speakers per episode is certainly an underestimate given that most RPGs have a DM and

Speaker 1	Now, let's meet the cast of DPR live who are often referred to as the tide 5. Brachii, the Goliath with arms Made of Stone his Barbarian rage will cut to the Bone. Ale and the Angelic Sorcerer of Helm he'll kill you with charm. It is known through the realm. Snatch is a halfling, you know, like a hobbit from Lord of the Rings and apple munching klepto. He'll steal all of your things. Pandora's a war priest who fights with goodness and piety. Unfortunately, he also fights with a crippling case of anxiety. O'ree Keyes. That's me. Wait, that's me
Speaker 4	a real devil with a
Speaker 1	lute with courage and Valor and an eye patch to boot. Alrighty, let's do this. It's time to roll dice. Let's all get Brody. It stays people who
Speaker 4	live. Nice
Speaker 3	nice. Hello everyone. Welcome. This is Dice paper roll live. Thank you all for coming. My name is Emil and I play brachii the Goliath Barbarian.
Speaker 1	Hi everybody. My name is Greg. I'm playing a land ay-ay-ron a on an AC Mars Sorcerer of Helm. Hey folks, my name is Ben and I play snatch of the halfling Rogue Uh, I'm Jack I play a re Keys song and I'm a tiefling Bard. I'm real nervous. There's a lot of people here today.
Speaker 2	And I'm Dan I play thund all the human fighter and cleric of Tempest. But I am also the dungeon master

Table 3: Example of an automatic transcript of the "Dice Paper Role" actual play podcast

Actual Play Podcast Corpus	
Number of Shows	555
Number of Episodes	38,713
Number of Episodes (transcribed)	32,602
Average no. of speakers per ep.	2.23
Average no. of turns per episode	597
Average no. of words per episode	11,374
Total no. of turns	19,478,136
Total no. of words	370,824,073

Table 4: Statistics for the actual play podcast corpus that we constructed

Play-By-Post Corpus	
Number of campaigns	896
Average players per campaign	8
Average turns per campaign	910
Average words per campaign	64,941
Total turns	815,106
Total words	58,187,526
Average dice rolls per campaign	594
Average annotations per campaign	94
Total dice rolls	532,270
Total annotations	84,447

Table 5: Statistics for the play-by-post corpus that we constructed

multiple players). Although podcast data is an interesting and plentiful source of game-play data, we ultimately decided to use text-based game play data instead.

3.2 Play-by-Post Data

Another source of game-play data are internet discussion forums where users engage in play-by-post of D&D and other roleplaying games. Figure 2 shows an example of part of the gameplay from the "play-by-post" forums from D&D Beyond.[1] These forums contain conversations similar to actual play podcasts, but players take turns writing a forum post describing their play instead of recording audio. Diarization and transcription errors are not a concern in play-by-post data, since it starts as text instead of audio.

The play-by-post data also has partial annotations. D&D Beyond provides a mechanism in its forum to roll dice using a "roll" tag. Their dice roller allows players to conduct the rolls that are used for D&D ability checks and in combat. We are able to locate the span of each post that corresponds to a dice roll, and extract information about the roll like what kind of die was used and what the total was.

Table 5 below summarizes the amount of play-by-post data that we collected from the D&D Beyond website (with permission from D&D Beyond).

We designed a set of rule-based heuristics to extract game state information from the play by post. These were implemented using regular expressions and NLP tools like named entity recognizers (Gardner et al., 2018). Although this heuristically extracted information is not perfect, it provides a reasonable approximation of the game state. It is useful for testing whether large language models can benefit from inclusion of complex state information for next utterance prediction and whether LLMs can be used for state tracking. We designed

[1] https://www.dndbeyond.com/forums/d-d-beyond-general/play-by-post

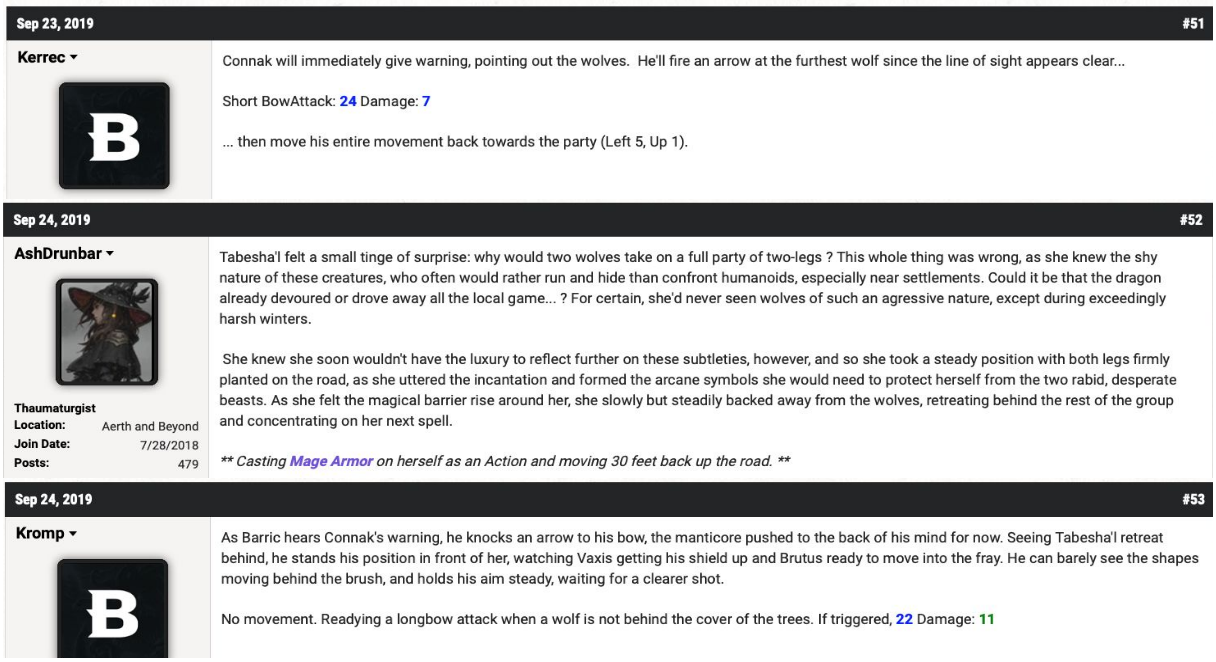

Figure 1: Example of 3 turns on the D&D Beyond play-by-post forum

rules to extract state information relating to character properties, combat and player actions.

Character properties

- Name: perform NER on all the player's turns in a campaign. The character's name is assigned to be the player's most frequently mentioned name, on the assumption that they tend to describe their own character's actions.
- Class: count how many times each D&D class is mentioned by each player. Most frequently mentioned class is their character's class.
- Race: On a player's first turn, check whether any of the D&D races are mentioned. Assign it to character. If not, apply other heuristics to guess it.
- Pronouns: Count pronouns mentioned by a player. Assign their character's pronouns to be the most frequent pronouns used by the player.
- Inventory: Regex that matches items occurring after character's personal pronouns (e.g. her sword).
- Spells known: Regex that matches cast followed by a spell name

The DM is assumed to be the player who has the first post in the game. The DM's entries in the dataset are scrubbed of other character properties, since they play multiple NPCs and monsters.

Combat

- We detect the start of combat when there is a roll for initiative, or when there are attack rolls before initiative (which happens in surprise attacks).
- Combat continues while there are attack rolls happening.
- Combat concludes after there are no rolls for a number of turns.
- In a combat span, we extract a list of monsters mentioned, and heuristically guess the number of each kind of monster.

Combat

- Dice rolls are marked in D&D Beyond posts. We detect the associated actions based on the kind of die used (D20 = a check, other dice are used for calculating damage if an attack check is successful)
- We use a regex to match the nearest pattern, which includes attack or a list of abilities like *acrobatics, animal handling, arcana, athletics,* etc.

- Damage rolls are matched with *damage, dmg, cure, heal, healing, points*.

These heuristics help to obtain at least one of the control features for around 58% of all conversational turns. We train a convolutional neural network (CNN) classifier using these conversational turns to predict all of the above control features for each conversational turn in training data.

The CNN classifier only uses current post text as input (no context). Table 6 estimates gives an estimate of the CNN's performance on filling in the state variables where the rule-based heuristic did not extract a value.

3.3 In-Character Versus Out-Of-Character Text

In addition to labeling the game states in our Play by Post data, we also labeled the text of each turn as being either spoken in-character (IC) or out-of-character (OOC). To do so, we crawled another Play by Post forum hosted at Giant in the Playground[2], where play happens on two discussion boards – one in-character and one out-of-characters. In the IC board, they also mask OOC actions with spoiler tags. We train a classifier to predict IC versus OOC text, and then apply it to all text in our D&D Beyond datasets.

4 Models

For our large language model, we use a 64B parameter version of Google's LaMDA language model (Thoppilan et al., 2022), which was trained on conversations. LaMDA is similar to other Transformer-based pre-trained language models like GPT-3. As with other pre-trained language models (Howard and Ruder, 2018), LaMDA can be fine-tuned to different tasks. The two tasks that we fine-tune LaMDA to perform are game state tracking and response generation. In both cases, the LLM can be thought of as a function that maps inputs onto an output. For instance, game state tracking is a language understanding task where the function takes in inputs like f(current utterance, previous state, history) $\rightarrow$ new state, and response generation is a language generation task where f(current state, history) $\rightarrow$ next utterance. The LLM functions are trained via the fine tuning process.

In our experiments we try a variety of different inputs to our LLM functions to see how they enable better learning of the tasks. We train our LLMs on the conversation history (which is typical in dialog modeling) and we also augment the conversations by conditioning other explicit signals. These conditioning signals can be thought of as sophisticated "control features", inspired by the CTRL language model (Keskar et al., 2019). During training, the model learns a relationship between the control features and appropriate responses. In turn, during inference, one can explicitly influence dimensions of the conversation – enabling more compelling dialogue – by setting the values of control features. These control features can be set dynamically, without necessitating fine-tuning or additional post-processing. Table 7 describes the control features we have proposed and describes how they could steer generation.

4.1 Baseline Pre-Training Data

LaMDA is trained on turn-based conversational data. For a conversation of length n, LaMDA takes as input the first $n-1$ turns, and the nth turn as the target. The example below illustrates the input and target for a three-turn conversation.

```
INPUT:
    TURN 1: I like the ocean
    TURN 2: Why?

TARGET:
    TURN 3: It's beautiful, vast, and mysterious.

TASK:
    TURN_1 SEP TURN_2 SEP ... SEP TURN_N-1 → TURN_N
```

4.2 D&D Fine-Tuning Data

Here is an example of the data used in our versions of LaMDA that are fine-tuned to on our D&D data.

TURN 1:

Text	You attack. You launch some fire onto the goblin closest to the wagon. And with that, he looks like he is on death's door. And the other goblin that you can see, the one that's not in the brush somewhere, just sort of stops in his tracks. What do you do next?
Player ID	0
Character	Dungeon Master
Race	N/A
Class	Dungeon Master
Pronouns	N/A
Inventory	N/A
In combat?	Yes
Action	Unknown

[2]https://forums.giantitp.com/forumdisplay.php?3-Play-by-Post-Games

State variable	Model	Type	Multi-valued	Availability	Evaluation metric	Performance
Character	Span labeller	Text	No	42%	-	-
Race	Classifier	Text	No	>58%	Macro AUC	0.45
Class	Classifier	Text	No	>75%	Macro AUC	0.71
Gender	Classifier	Text	No	42%	Macro AUC	0.92
Inventory	Span labeller	Text	Yes	11%	-	-
In combat?	Classifier	Score	No	100%	Accuracy	0.91
Action	Classifier	Text	Yes	20%	Macro AUC	0.92

Table 6: The estimated performance of our CNN classifier on predicting state values for turns where our rule-based heuristics did not predict a value

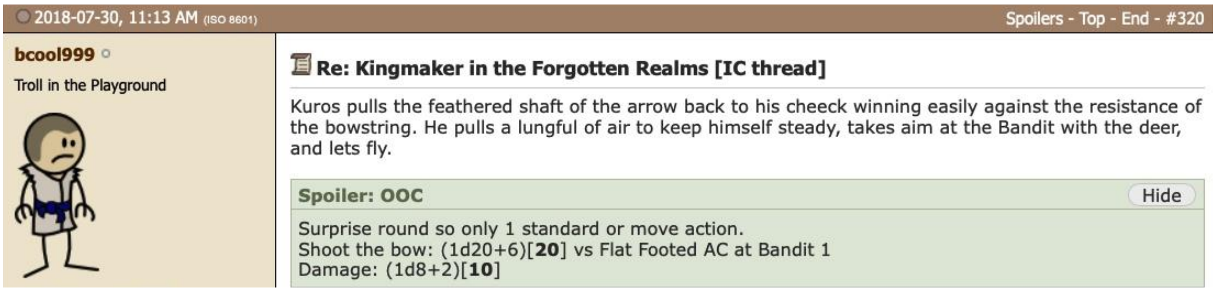

Figure 2: An example post from the Giant in the Playground forums where the text is segmented into in-character (IC) and out-of-character (OOC) portions.

Control Feature	Description	Expected Impact on Model's Output
Player ID	Player writing a given dialog turn	Connects the current turn to the player's previous turns, which is important in multi-party conversations.
IC versus OOC	Whether a player is in-character or out-of-character for a given dialog turn	Changes whether the generated text is more like descriptive text found in a novel, or more like a discussion of rules and strategies.
Character Name	Name of the character being played by the player of a given dialog turn	IC descriptions use the character's name.
Character Class	D&D classes[3]	Character classes perform different actions (e.g. wizards cast spells, thieves pick locks)
Character Race	D&D fantasy races[4]	Different physical characteristics (e.g. halflings are small, dragonborn have scales).
Character Pronoun	The character's pronouns	Uses the correct pronouns when describing the character.
Character Actions	List of actions taken by the character in the current turn	Allows a description to be generated for a given action. The action can be thought of as a goal for the description.
Combat	Whether the players are currently engaged in combat or not during a given dialog turn	Affects the likelihood of actions (e.g. attacks are more likely during combat and investigations checks are more likely outside of combat)

Table 7: Our LLMs are conditioned on a variety of control features that allow the models to better learn what kind of text to generate for the next utterance prediction task

TURN 2:

Text	I grab my axe and bring it down on the wounded goblin.
Player ID	1
Character	Magnus
Race	Human
Class	Fighter
Pronouns	he/him
Inventory	Axe
In combat?	Yes
Action	Attack
Action	Unknown

4.3 Next Utterance Prediction Models

LLM-Dialog We call our baseline model LLM-Dialog. It is a LaMDA dialogue model that does not use not use any D&D data.

LLM-DND LLM-Dialog that has been finetuned on Play-by-post D&D gameplay dataset using *no* control features
TASK:
TURN_1 SEP TURN_2 SEP ... SEP $\text{TURN}_{N-1} \rightarrow \text{TURN}_N$
where TURN_i denotes text along with player id for ith dialog turn and SEP is the separator. Note that the last turn TURN_N is the target and previous conversation history is the input for this task.

LLM-DND-ALL-CTRL LLM-Dialog that has been finetuned on Play-by-post D&D gameplay dataset using control features for *all* dialog turns upto the last or target turn.
TASK:
TURN_1 STATE_1 SEP TURN_2 STATE_2 SEP ... SEP TURN_{N-1} STATE_{N-1} SEP $\text{STATE}_N \rightarrow \text{TURN}_N$
where STATE_i denotes the set of control features for ith dialog turn, e.g., `in-combat : 0 | class : fighter | character : magnus | race : human | pronouns : he/him | inventory : axe | actions : attack`

LLM-DND-PREV-CTRL LLM-Dialog that has been finetuned on Play-by-post D&D gameplay dataset using control features for all *previous* dialog turns to the last or target turn.
TASK:
TURN_1 STATE_1 SEP TURN_2 STATE_2 SEP ... SEP TURN_{N-1} $\text{STATE}_{N-1} \rightarrow \text{TURN}_N$

LLM-DND-RECENT-CTRL LLM-Dialog that has been finetuned on Play-by-post D&D gameplay dataset using control features for most **recent** dialog turn before the last or target turn.

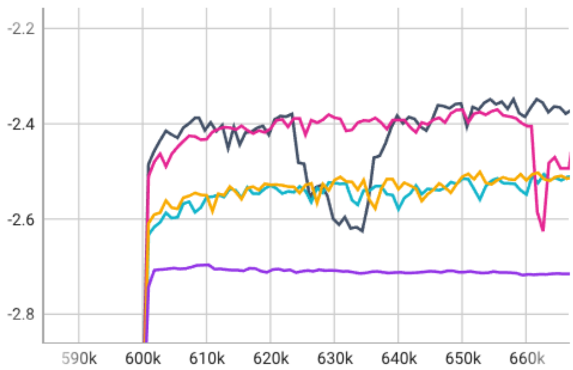

Figure 3: Perpelxity of our models after pretraining on generic dialogue data for 600k steps, and then finetuning to our data for a further 60k steps

Model	Perplexity	Token accuracy
LLM-Dialog	2.65	44.61
LLM-DND	2.50	46.92
LLM-DND-PREV-CTRL	2.51	46.84
LLM-DND-RECENT-CTRL	**2.34**	**49.67**
LLM-DND-ALL-CTRL	2.37	49.02
LLM-DND-Podcast	2.53	46.64
LLM-Podcast	2.69	44.6

Table 8: Perpelxity and token accuracy of our models after finetuning to our data

TASK:
TURN_1 STATE_1 SEP TURN_2 STATE_2 SEP ... SEP TURN_{N-1} $\text{STATE}_{N-1} \rightarrow \text{TURN}_N$

LLM-Podcast LLM-Dialog that has been finetuned on (transcribed) Dungeons & Dragons podcasts using **no** control features

LLM-DND-Podcast LLM-Dialog that has been finetuned on both (transcribed) Dungeons & Dragons podcasts and using Play-by-post D&D gameplay dataset **no** control features

4.4 Dev set perplexity during training

Each of our models starts from a pretrained LaMDA model trained for 600K steps and then is finetuned for a further 60K steps. Figure 8 plots the Negative log perplexity on our development set, and Table **??** shows the final perpexity and token accuracies on the dev set. At the end of finetuning, the models with the best perplexity scores and the best token accuracy scores were LLM-DND-RECENT-CTRL and LLM-DND-ALL-CTRL.

5 Manual Evaluation

To evaluate the quality of our models for the task of next utterance prediction in D&D , we perform a human evaluation. We recruited professional raters to perform a manual evaluation. They read a version of the content that was provided to the models – the seven turns of conversational history plus a list of players and the names/classes of the characters that they played. Then they were shown several model outputs for the context (or the "gold", which was the actual next turn in the game), The annotators asked to rate each output along the three dimensions, following the evaluation procedure used for the Meena LM (Adiwardana et al., 2020):

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

The full annotator instructions are given in the Appendix. A mock up of the user interface is given in Figure 4.

5.1 Raters

Because of the specialized nature of the D&D domain, we recruited 6 professional raters rather than crowd workers to perform the task. The raters were selected based on their professed interest in the fantasy genre, and on their background with D&D . All raters were fantasy fans, and 5 of the 6 had played D&D . 3 raters had been DM in a game before.

5.2 Agreement

Our raters annotated 500 system outputs with 3-way redundancy on each output. For the binary sense and specific scores, the pairwise annotator agreement was 0.8, with a chance-adjusted Radolph Kappa score of 0.6. For the scalar interestingness scores, the Kendall's Tau correlation was 0.46.

5.3 Model comparison

Table 9 shows the average sense, specific and interestingness scores for the systems, and for the human-written gold response. All of the D&D adapted systems outperform the vanilla dialogue system. The added control features do not seem to differ substantially from the LLM that is adapted to the D&D data without any control features.

We further analyzed the systems to see if the in-character versus out-of-character turns made a difference. Table 10 shows that interestingness substantially increased on in-character turns than when the output was generated out-of-character. Because our CTRL models allowed the system to intentionally generate in-character responses, it was able to intentionally produce in-character responses, resulting in substantially improved interestingness scores for in-character turns.

Table 11 shows a head-to-head comparison of systems based on what fraction of time one system's interestingness score was higher than another system's. In this analysis we found that the systems with the highest average head-to-head comparison were all of the systems that used the CTRL variables.

5.4 Qualitative Example

Table 12 shows example outputs from different versions of our LLMs. We gave the models the final seven turns in the game play example –from "With his eagle eyes, Magnus spots two dead horses. . . " to "... What do you do next?", and then generate what each model thinks the next utterance ought to be. The vanilla dialogue system with no D&D data generates an irrelevant reply about Discord. The dialogue system fine tuned on D&D data without state generates a relevant response. The dialgoue system that adds conditioning control features allows the model to generate specific replies for each of the different characters. These replies are relevant to the character classes – the wizard casts a spell, and the Dwarf cleric shouts the name of a Dwarf god.

6 Game State Tracking Model

We conducted one final pilot experiment to try to evaluate if a LLM could be fine-tuned to perform game state tracking for D&D using our heuristically annotated partial game state features. We trained LLM-DND-GST (Game State Tracking). It is a LLM-Dialog that has been finetuned on Play-by-post D&D gameplay dataset using control features or state variables for all **previous** dialog turns to the last turn (and all dialog turns including the last turn) to predict the dialog state of the last turn.
TASK:
TURN_1 STATE_1 SEP TURN_2 STATE_2 SEP

Context	**Player 10 - Del**	Ratings
Player 10 - Del (Barbarian) Player 0 - Dungeon Master Player 12 - Aramis (Bard) Player 8 - Xavier (Bard) -- Player 10 - Del (Barbarian): Del gives a final kick to the dead orc that he just killed then proceeds to hunt through the main room then the store room, looking for his axe and other gear Player 0 - Dungeon Master : Del - make an investigation check. Player 10 - Del (Barbarian): 3 Player 0 - Dungeon Master : The barbarian begins tossing things around, overturning tables and tearing crates and sacks open, though his still simmering rage at being locked up prevents him from undertaking a thorough search. He turns up a shortbow, a light crossbow, a few javelins, 2 quivers of arrows, 2 cases of crossbow bolts, a crate full of mismatched leather armour and a few sacks full of preveserved fruit and vegetables. There is no sign of the equipment that was taken from him last night. Player 12 - Aramis (Bard) : ""anyone seen any keys? or my stuff?"" aramis calls out to everyone Player 8 - Xavier (Bard) : *quickly picking up the dead orcs cleaver so a better weapon Xavier turns to the rest of the party* ""fighting is not over, we are not sure whats to be found up there, raid the store room, we must be ready for more combat"" *i turn to del* ""would you stand with me to defend the rest while they re-arm?"" *i offer the crowbar to del* *then turn back to the party* ""and if you find a shield and sword in there, i require a stronger defence and fist""	Response 1 **Del quickly searches the store room, grabbing whatever weapons he find useful. He then heads back to the others "I found another crossbow and arrows, and a dagger. I have a short bow, but I'll leave it here if someone else wants it. I also found one of my javelins that I was missing."**	Does this response make sense? ❏ Makes sense ❏ Does not make sense Is this response specific to the given context? ❏ Specific ❏ Not specific How interesting is this response on a scale of 10? ____
	Response 2 **del takes the crowbar and tells xavier to find hes ****en axe and ill keep good care of him**	Does this response make sense? ❏ Makes sense ❏ Does not make sense Is this response specific to the given context? ❏ Specific ❏ Not specific How interesting is this response on a scale of 10? ____
	Response 3 **Anyone see my axe, and leather?**	Does this response make sense? ❏ Makes sense ❏ Does not make sense Is this response specific to the given context? ❏ Specific ❏ Not specific How interesting is this response on a scale of 10? ____
	Response 4 **Noticing Del is searching the same room. Del calls out to him "X, I just killed an orc you could use its stuff if you like."**	Does this response make sense? ❏ Makes sense ❏ Does not make sense Is this response specific to the given context? ❏ Specific ❏ Not specific How interesting is this response on a scale of 10? ____

Figure 4: The user interface that our raters used to evaluate the quality of our model's next utterance prediction.

	Sense	**Specific**	**Avg. Interesting**	**Avg. Interesting Rank**
LLM-Dialog	0.81	0.85	3.57	3.79
LLM-DND	0.9	0.9	3.91	3.46
LLM-DND-ALL-CTRL	0.87	0.88	3.92	3.43
LLM-DND-PREV-CTRL	0.86	0.88	3.96	3.35
LLM-DND-RECENT-CTRL	0.88	0.9	3.96	3.46
GOLD	0.92	0.92	4.17	3.11

Table 9: Average sense, specific and interestingness scores for our systems, and for human-written gold responses

In-char/Out-char	**Sense**	**Specific**	**Avg. Interesting**	**Avg. Interesting Rank**
LLM-Dialog	0.80/0.81	0.84/0.85	3.60/3.54	4.10/3.41
LLM-DND	0.89/0.91	0.92/0.89	4.09/3.69	3.59/3.29
LLM-DND-ALL-CTRL	0.90/0.83	0.91/0.85	4.28/3.47	3.28/3.62
LLM-DND-PREV-CTRL	0.87/0.85	0.89/0.87	4.23/3.63	3.33/3.37
LLM-DND-RECENT-CTRL	0.91/0.85	0.93/0.87	4.38/3.45	3.10/3.63
GOLD	0.95/0.88	0.94/0.89	4.63/3.61	2.84/3.44

Table 10: Analysis of in-character (IC) versus out-of-character (OOC) turns. Numbers are IC/OOC scores.

Test	Dialog	DND	DND-ALL-CTRL	DND-PREV-CTRL	DND-RECENT-CTRL	GOLD	Average	Ranked
Dialog	-	0.37	0.37	0.37	0.37	0.31	0.36	6
DND	0.63	-	0.46	0.47	0.45	0.40	0.48	5
DND-ALL-CTRL	0.63	0.54	-	0.47	0.47	0.40	0.50	4
DND-PREV-CTRL	0.63	0.53	0.53	-	0.50	0.44	0.52	3
DND-RECENT-CTRL	0.63	0.53	0.53	0.50	-	0.44	0.53	2
GOLD	0.69	0.60	0.60	0.56	0.56	-	0.60	1

Table 11: Head-to-head comparison of systems of how often one system's interestingness score was higher than anothers

Model	Output
LLM-Dialog	moon_monk you can just join our discord conversation. it's free and easier to deal with
LLM-DND	I cast Ice Knife (I'm assuming this is not one of the ones I can't do?)
LLM-DND-PREV-CTRL Responding as Magnus (human fighter, inventory: axe) In combat	I turn from the goblin and go to help Merle and Taako with getting this wagon moving again. Magnus can't see them anymore with such a dense forest. They're likely in a battle right now... I have to find them as fast as I can!
LLM-DND-PREV-CTRL Responding as Taako (elf wizard, inventory: spellbook) In combat	If the goblins still don't look like they're getting closer, Taako will yell, "Alright you little monsters! Let's see what ya got!" And he'll cast Chill Touch on one of the goblins that isn't dead/on its doorstep. Attack: 22 Damage: 6
LLM-DND-PREV-CTRL Responding as Mearle (dwarf cleric, inventory: holy symbol) In combat	Merle steps away from the wagon. He draws his sword and takes aim at the hidden goblin! Attack: 27 Damage: 8 He puts his free hand on the pommel of his sword and shouts his battle cry, "For Clanggeddin!" His eyes glow brightly red as he strikes!

Table 12: Example outputs our models for the next utterance predicted, given the sample gameplay given in Table 2 as input.

State variable	Majority	LLM-DND-GST
All	72.45	81.48
Combat	88.46	82.41
Character Class	58.24	76.37
Character Name	58.24	76.92
Character Race	74.73	78.57
Character Pronouns	58.24	86.81
Character Actions	80.22	84.61

Table 13: Average accuracy for our Game State Tracking LLM on the slot-filling our state variables, compared to a majority class baseline.

... SEP TURN$_{N-1}$ STATE$_{N-1}$ TURN$)N \rightarrow$ STATE$_N$

We analyzed the accuracy of the LLM-DND-GST model its ability to do slot-filling for each of the individual game states, and compared its performance to a simple baseline that always output the the majority class. The results are shown in Table 13. The average accuracy of the dialogue state tracker is better than the majority class baseline, but likely falls short of being useful when it comes to joint accuracy. The joint accuracy for the Majority class baseline is 58.24 and the joint accuracy for LLM-DND-GST is 58.24. This suggests that accurately tracking the state of the game may require additional machinery beyond a fine-tuned LLM.

7 Limitations

One limitation of our human evaluation is that it is a static evaluation. The raters are simply reading the outputs of the model, and there is no interactive evaluation wherein they engage in gameplay with the system. An interactive user-study would be required before any claims could be made about how well AI is able to play D&D alongside human players.

Because our state information was created heuristically, and therefore it potentially contains errors. It is also incomplete. There are several kinds of state tracking variables that would be useful to include, but were not possible to heuristically extract from our data. These include:

- Current HIT points totals for each character and monster. This is perhaps the most relevant state that we're missing.
- Slot-filler values for attacks
  1. Attacker
  2. Target
  3. Weapon
  4. Damage amount
- For a check roll, was the check successful or not? This depends on the armor class of the target of an attack, or the difficulty class of a skill (often hidden). Therefore, although we have dice roll totals, we do not know if the attempt succeeded or failed.

To address this problem in the future, we have begun a collaboration with the developer of Avrae[5], which is a Discord bot for playing D&D online. Avrae contains many of the state variables that are missing from our current annotations.

8 Related Work

Although there has been work done on text-based game playing (Haroush et al., 2018; Yao et al., 2020; Dambekodi et al., 2020), these games still provide points for taking correct actions and generally have a limited vocabulary to work with. Creating text games (Ammanabrolu et al., 2020a; Fan et al., 2020) is more challenging but closer to the type of world-creating job the DM has in D&D . There has also been work on persona/character generation (Louis and Sutton, 2018; Prabhumoye et al., 2019) in stories, sometimes within D&D itself (Louis and Sutton, 2018). Others (Urbanek et al., 2019; Ammanabrolu et al., 2020b) have realized that non-player characters (NPCs) are lacking in their abilities to speak and act in text games.

Thankfully, we do not have to start from scratch when working toward the vast challenge of creating a D&D player and can rely upon the findings of the automated story generation community. Neural language models have become increasingly more popular for story generation (Roemmele, 2018; Martin et al., 2018a; Mathewson et al., 2019; Hou et al., 2019). We have also started to see storytelling with transformers (See et al., 2019; Peng et al., 2021; Branch et al., 2021). Transformer-based storytelling systems have even been introduced to the general public thanks to the popularity of AI Dungeon (Walton, 2019). Although neural networks possess a lot of power in terms of what text they generate, they are still limited in their ability to produce longer spans of coherent text.

[5]https://avrae.io

Many (Fan et al., 2018; Yao et al., 2019; Ippolito et al., 2019; Tambwekar et al., 2019; Ammanabrolu et al., 2020b; Rashkin et al., 2020) have improved the coherence of neural storytellers by splitting the generation into two steps: ideation of the story plot, followed by the realization of sentences. This *controllable story generation* is where we are currently seeing a lot of work within neural automated story generation since it enables the neural network(s) to focus on a single task at a time.

Due to the conversational nature of D&D , we decided to use a dialog-based system. Deep neural networks have been used for dialog agents for a while (Serban et al., 2016), with a shift toward using transformers in recent years (Zhang et al., 2019; Ghazarian et al., 2021). Like in automated story generation and other neural text generation tasks, we are also seeing controllability being an important factor being integrated into systems. This includes using deep reinforcement learning techniques to guide the dialog toward a goal (Li et al., 2016; Saleh et al., 2020) or controlling for style (Zhang et al., 2018; Smith et al., 2020).

In this paper, we use LaMDA, a transformer-based open-domain dialogue system that building on the Meena model (Adiwardana et al., 2020). The original Meena model was an end-to-end model trained on public conversations found on social media. Controllable text generation with transformers has been seen before with CTRL (Keskar et al., 2019), a language model that is conditioned on a given "control code" in addition to the Bayesian history. This work takes a similar approach. We integrate contextual information such as character descriptions, actions, and in- and out-of-character classifications.

We have finetuned our LaMDA models on data crawled from D&D Beyond[6]. This data contains both in-character and out-of-character dialog and can be used in conjunction with Rameshkumar and Bailey (2020)'s dataset from Critical Role (a D&D podcast), Louis and Sutton (2018)'s dataset from roleplayerguild.com (a D&D forum), and/or Urbanek et al. (2019)'s crowdsourced LIGHT dataset. For the purposes of this work, we will only be working with our D&D Beyond dataset.

[6]https://www.dndbeyond.com/

9 Discussion and Conclusions

We find that training on D&D data results in much higher quality outputs than a vanilla dialogue system (as expected), that controlling the model to generate in-character responses results in substantially more interesting output, and that conditioning on game state information qualitatively results in responses that are appropriate to the character class. Our preliminary experiments with using the large language models to perform game state tracking show low performance even after fine-tuning, suggesting that other models may be required for an AI to play D&D track the full state of the game.

10 Acknowledgments

We would like to thank Antony Pegg of Fandom for granting us permission to use D&D Beyond's forum data for research purposes.

Thank you to Rich Burlew and forum moderator truemane for granting us permission to crawl the Giant in the Playground forum and to build models using the forum posts.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu Quoc, and V Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.

Prithviraj Ammanabrolu, William Broniec, Alex Mueller, Jeremy Paul, and Mark O. Riedl. 2020a. Toward Automated Quest Generation in Text-Adventure Games. In *International Conference on Computational Creativity (ICCC)*.

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. 2020b. How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds. In *The North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Boyd Branch, Piotr Mirowski, and Kory Mathewson. 2021. Collaborative Storytelling with Human Actors and AI Narrators. In *International Conference on Computational Creativity (ICCC)*, pages 97–101.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sahith Dambekodi, Spencer Frazier, Prithviraj Ammanabrolu, and Mark O Riedl. 2020. Playing Text-Based Games with Common Sense. In *Wordplay: When Language Meets Games Workshop (NeurIPS 2020)*.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898.

Angela Fan, Jack Urbanek, Pratik Ringshia, Emily Dinan, Emma Qian, Siddharth Karamcheti, Shrimai Prabhumoye, Douwe Kiela, Tim Rocktäschel, Arthur Szlam, and Jason Weston. 2020. Generating Interactive Worlds with Text. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1693–1700.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Sarik Ghazarian, Zixi Liu, Tuhin Chakrabarty, Xuezhe Ma, Aram Galstyan, and Nanyun Peng. 2021. DiSCoL: Toward Engaging Dialogue Systems through Conversational Line Guided Response Generation. *arXiv preprint arXiv:2102.02191*.

Matan Haroush, Tom Zahavy, Daniel J. Mankowitz, and Shie Mannor. 2018. Learning How Not to Act in Text-Based Games. In *Workshop Track at ICLR 2018*, pages 1–4.

Chenglong Hou, Chensong Zhou, Kun Zhou, Jinan Sun, and Sisi Xuanyuan. 2019. A survey of deep learning applied to story generation. In *Smart Computing and Communication*, pages 1–10. Springer International Publishing.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Daphne Ippolito, David Grangier, Chris Callison-Burch, and Douglas Eck. 2019. Unsupervised Hierarchical Story Infilling. In *Proceedings of the First Workshop on Narrative Understanding*, pages 37–43.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 633–644.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Jiwei Li, Will Monroe, Alan Ritter, and Dan Jurafsky. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Annie Louis and Charles Sutton. 2018. Deep Dungeons and Dragons: Learning Character-Action Interactions from Role-Playing Game Transcripts. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 708–713.

Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. 2018a. Event Representations for Automated Story Generation with Deep Neural Nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 868–875, New Orleans, Louisiana.

Lara J. Martin, Srijan Sood, and Mark O. Riedl. 2018b. 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.

Kory W. Mathewson, Pablo Samuel Castro, Colin Cherry, George Foster, and Marc G. Bellemare. 2019. Shaping the Narrative Arc: An Information-Theoretic Approach to Collaborative Dialogue. *arXiv preprint arXiv:1901.11528*, page 20.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Johannes E. M. Mosig, Vladimir Vlasov, and Alan Nichol. 2020. Where is the context? - A critique of recent dialogue datasets. *CoRR*, abs/2004.10473.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2021. Inferring the Reader: Guiding Automated Story Generation with Commonsense Reasoning. In *Workshop on Narrative Understanding at NAACL-HLT 2021*.

Shrimai Prabhumoye, Khyathi Raghavi Chandu, Ruslan Salakhutdinov, and Alan W Black. 2019. "My Way of Telling a Story": Persona based Grounded Story Generation. In *Storytelling Workshop at ACL 2019*, pages 11–21.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5121–5134.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295.

Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting Extraneous Content in Podcasts. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 1166–1173, Online. Association for Computational Linguistics.

Melissa Roemmele. 2018. *Neural Networks for Narrative Continuation*. Ph.D. thesis, University of Southern California.

Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind Picard. 2020. Hierarchical Reinforcement Learning for Open-Domain Dialog. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 8741–8748.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do Massively Pretrained Language Models Make Better Storytellers? In *Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. In *Conference on Neural Information Processing Systems (NeurIPS), Workshop on Learning Methods for Dialogue*, pages 1–6.

Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2020. Controlling Style in Generated Dialogue. *arXiv preprint arXiv:2009.10855*.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2019. Controllable Neural Story Plot Generation via Reinforcement Learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, Macau, China.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *CoRR*, abs/2201.08239.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to Speak and Act in a Fantasy Text Adventure Game. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Nick Walton. 2019. AI Dungeon.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-And-Write: Towards Better Automatic Storytelling. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 7378–7385.

Shunyu Yao, Rohan Rao, Matthew Hausknecht, and Karthik Narasimhan. 2020. Keep CALM and Explore: Language Models for Action Generation in Text-based Games. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8736–8754.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213, Melbourne, Australia.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv preprint arXiv:1911.00536*.

A Lost Mine of Phadelver Adventure

Here is an excerpt from the adventure book that the Dungeon Master was using in our example game play. The adventure book provides boxed text, which is descriptive text to be read aloud to paraphrased. It also gives details about the combat that is about to ensure, and links to relevant game rules (like stealth checks, and statistics about the monsters that the characters will be in combat with).

The adventure begins as the player characters are escorting a wagon full of provisions and supplies from Neverwinter to Phandalin. The journey takes them south along the High Road to the Triboar Trail, which heads east (as shown on the overland map). When they're a half-day's march from Phandalin, they run into trouble with goblin raiders from the Cragmaw tribe.

Read the boxed text when you're ready to start. If you create a different adventure hook, skip to the second paragraph and adjust the details as necessary, ignoring the information about driving the wagon.

> In the city of Neverwinter, a dwarf named Gundren Rockseeker asked you to bring a wagonload of provisions to the rough-and-tumble settlement of Phandalin, a couple of days' travel southeast of the city. Gundren was clearly excited and more than a little secretive about his reasons for the trip, saying only that he and his brothers had found "something big," and that he'd pay you ten gold pieces each for escorting his supplies safely to Barthen's Provisions, a trading post in Phandalin. He then set out ahead of you on horse, along with a warrior escort named Sildar Hallwinter, claiming he needed to arrive early to "take care of business."
>
> You've spent the last few days following the High Road south from Neverwinter, and you've just recently veered east along the Triboar Trail. You've encountered no trouble so far, but this territory can be dangerous. Bandits and outlaws have been known to lurk along the trail.
>
> You've been on the Triboar Trail for about half a day. As you come around a bend, you spot two dead horses sprawled about fifty feet ahead of you, blocking the path. Each has several black-feathered arrows sticking out of it. The woods press close to the trail here, with a steep embankment and dense thickets on either side.

Four goblins are hiding in the woods, two on each side of the road. They wait until someone approaches the bodies and then attack.

This will likely be the first of many combat encounters in the adventure. Here are the steps you should follow to run it effectively:

- Review the goblin stat block in appendix B. Since the goblins are hiding, you'll need to know their Stealth skill modifier: +6.
- Check to see who, if anyone, is surprised. The party can't surprise the goblins, but the goblins might surprise some or all of the characters. Make a Dexterity (Stealth) check for the goblins: roll one d20 for all of them, add their Stealth skill modifier (+6) to the roll, and compare the total to the characters' passive Wisdom (Perception) scores. A character whose score is lower than the goblins' check total is surprised and therefore can't do anything on his or her first turn in the combat (see "Surprise" in the Basic Rules).
- Use the initiative rules in the Basic Rules to determine who acts first, second, third, and so on. Keep track of everyone's initiative count on a piece of paper.
- When the time comes for the goblins to act, two of them rush forward and make melee attacks while two goblins stand 30 feet away from the party and make ranged attacks. The goblins' stat block contains the information you need to resolve these attacks. For more information on what the goblins can do on their turn, see chapter 9, "Combat," in the Basic Rules. When three goblins are defeated, the last goblin attempts to flee, heading for the goblin trail

B Annotation Guidelines

B.1 Annotation task

In this task, you will see part of a conversation between a few people playing D&D . The players and their characters are listed at the beginning of the conversation. The conversations that are shown as context are real conversations from players. Your job is to read the context and then rate different responses for a player/character given conversational context. Please note that the context you are given represents only a part of the players' past conversations/interactions with one another during the game.

For each response, you would be asked the following questions.

- Does the response make sense?
  - Use your common sense here. Is the response completely reasonable in terms of the rules of D&D ?
  - The response "makes sense" if it is cohesive as a standalone statement, consistent with the rules of the game, and the elements/entities mentioned are plausible, given the prior context.
  - If anything seems off—not fluent, confusing, illogical, out of context, or wrong according to the rules of D&D —then rate it as Does not make sense. If in doubt, choose Does not make sense.
- Is the response specific?
  - You may be asked to assess whether the response is specific to a given context. In other words, do you think that the response represents a good thing for the character to do now?
  - The response is "specific" if it flows logically from the narrative established by the prior context.
    - 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.
- How interesting is the response?
  - You may be asked to score the response for its interestingness on a scale of 10. Choose a high score for "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, or if you're unsure, then it is Less Interesting.

B.2 Survey of Raters

We recruited raters who had a background in role playing games and an understanding of the fantasy genre. We surveyed our raters, asking them the following questions:

1. Have you ever played Dungeons and Dragons or another role playing game before?
2. If so,
   - roughly how many times have you played
   - were you a player or a game master or both
3. If not,
   - what kind of exposure do you have to Dungeons and Dragons? (For example, have you seen it referred to in TV or movies)
4. Are you a fan of the fantasy genre (like Lord of the Rings)?

Our 6 raters responded to the survey as follows: 5 out of the 6 have played D&D or another role playing game before. All 5 of those who have played D&D /other role playing games before have played more than 6 times. Of the 5 who have played D&D /other role playing games before, 3 played as both Game Master and Player. For the one who had not played D&D /other role playing games, they indicated they had not had much exposure to D&D through TV or other channels. All 6 answered that they were fans of the fantasy genre.