

# Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence

Anonymous ACL submission

## 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 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 led 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. Video games like StarCraft or classic Atari games have provided test beds for reinforcement

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

learning where systems must learn game playing policies by interacting with dynamic worlds where the game state is not as easily repressed as board games. 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. Incorporating Neural LMs into a game 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) that denotes their abilities, and a fantasy race (like elf, dwarf, human). 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 Burnside), 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 roleplaying 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).

<b>Player (character)</b>	<b>Game Dialogue</b>	<b>D&amp;D Game Description and AI challenges</b>
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.	<i>This game is based on the D&amp;D starter adventure called "Lost Mine of Phandelver". 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.</i> AI challenges: <b>Generation of stories and descriptive text</b>
Griffin (DM)	Roll a perception check for me. Perception is a wisdom skill, so be sure to add your wisdom modifier.	<i>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.</i> AI challenges: <b>Knowledge base population (extraction of rules from a rulebook)</b>
Clint (out of character)	I got an eight.	<i>Clint has rolled his dice. The number is low so his character fails the check.</i> AI challenges: <b>Multi-party dialogue</b>
Justin (out of character)	I got a six.	<i>Justin also fails. Neither character sees the thing that requires the perception check.</i>
Travis (out of character)	I rolled a natural twenty plus my wisdom modifier is 23.	<i>Travis rolls high number and succeeds on the check.</i> AI challenges: <b>Understanding rules, determining success or failure</b>
Griffin (DM)	With his eagle eyes, Magnus spots two dead horses lying in the middle of the road about 200 feet ahead of you.	<i>The DM describes what happens as a result of the success.</i> AI challenges: <b>Reasoning about consequences of success or failure, descriptive text generation</b>
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.	<i>Travis is describing what he is doing using "in character" language.</i> AI challenges: <b>Persona-based chat</b>
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.	<i>The DM describes the start of a battle with several monsters.</i> AI challenges: <b>State tracking (in combat v. out of combat).</b>
Travis (out of character)	How many goblins are there?	AI challenges: <b>Question answering, state tracking (how many monsters).</b>
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: <b>Question answering, Descriptive text generation from game state.</b>
Clint (Merle)	I will cast sacred flame at the nearest one. If it fails a dexterity saving throw, it takes 6 points of damage.	<i>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.</i> AI challenges: <b>Intent detection (perform attack action against a particular goblin)</b>
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?	<i>The DM rolls for the monster, updates the state of its health meter, and describes the result of Merle's attack.</i> AI challenges: <b>Reasoning about rules, state tracking (monster's HIT points), descriptive text generation.</b>

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

157	• It covers a spectrum of task oriented and non-	• Character specific state variables (remain con-	204
158	task oriented (e.g. chit chat) dialog.	stant throughout the game). These include	205
159	• It has many participants in the conversation,	Player ID, Character’s name, pronouns, class	206
160	since there are several players in the game.	and fantasy race, Items in character’s inven-	207
161	• It conveys narrative elements including de-	tory.	208
162	scriptions of events, that denote changes in	• Combat related state variables (change period-	209
163	the state of the game.	ically) These include Whether the players are	210
164	We evaluate a state-of-the-art language model, to	engaged in combat, Monsters in combat	211
165	understand how conditioning generation on differ-	• Dialog turn level state variables. These are	212
166	ent elements improves the quality of the generated	actions related to dice rolls .	213
167	text.		
168	<b>2 Tasks</b>	<b>3 Datasets</b>	214
169	We trained a large language model (LLM) to per-	For this paper, we have created a novel dataset	215
170	form two tasks: <b>Next Utterance Prediction</b> and	for our dialogue-oriented test of AI’s ability to	216
171	<b>Game State Tracking</b> .	play Dungeons Dragons. We investigated two	217
172	<b>Next Utterance Prediction</b> We trained our lan-	sources of data of people playing the game. The	218
173	guage model on a corpus of human conversations	first source of data was “Actual Play” podcasts,	219
174	(see Datasets section) to predict the next utterance.	where people record themselves actually playing	220
175	We varied the conditioning information to examine	RPGs. The second Play-By-Post data scraped from	221
176	the effects on the quality of predicted next utterance.	a web forum where people play by taking turns	222
177	In all variations, we included the conversational his-	posting on the forum to describe their move. Ulti-	223
178	tory as input. We used most recent 7 conversational	mately, we focused primarily on the Play-By-Post	224
179	turns as input (in our example in the Table above,	data rather than the Actual Play Podcast data, since	225
180	that would be from Griffin saying “With his eag-	our automatic transcripts of the podcast audio had	226
181	le eyes...” up until “What do you do next?”).	some speaker diarization errors and misrecognized	227
182	Given the conversational input (and other input in	words. Furthermore, using data from podcasts is	228
183	the variant models), the LLM must generate the	not straightforward due to the inclusion of extrane-	229
184	next utterance, such that it is both interesting and a	ous content, such as advertisements (Reddy et al.,	230
185	plausible next turn in the D&D game.	2021).	231
186	<b>Game State Tracking</b> In this task, rather than	<b>3.1 Actual Play Podcasts</b>	232
187	outing the next utterance, we had the model pre-	Actual Play podcasts are a genre of podcasts where	233
188	dict the game state for a given dialogue turn in	people record themselves playing RPGs. We col-	234
189	the conversation. We have kept the state defini-	lected lists of popular actual play podcasts from	235
190	tion similar to task-oriented dialogue state tracking	the web, and expanded the list by searching for	236
191	(DST). In DST, the dialogue state is a collection	additional shows with relevant keywords including	237
192	of slot-value pairs. In our case, each slot is a state	Actual Play, Role Playing Game, D&D 5e. Ex-	238
193	variable feature related to D&D games. Our target	tracted all podcast episodes for these candidate	239
194	slot values do not need to appear as a word in the	podcast shows, and transcribed all podcast episodes	240
195	dialogue context. We track several game states as-	using Google Cloud Speech-to-Text API (discard-	241
196	pects including some that remain relatively static	ing about 15% of the episodes because of download	242
197	throughout the game (character attributes like their	errors or transcription failures). Table 3 shows a	243
198	pronouns, class, fantasy race, and their inventory),	fragment of one of the podcast transcript. Table	244
199	some that change periodically (like being in com-	4 summarizes the amount of Actual Play podcast	245
200	bat or out of combat), and some that change from	data that we collected.	246
201	turn to turn (like what action the player is taking).	Dealing with speech data presents several chal-	247
202	In this paper we attempt to track the following	lenges, including large audio files (the mean	248
203	game state variables:	episode length was 1 hour and 17 minutes, with a	249
		max of 7 hours and 23 minutes), automatic speech	250
		recognition errors, and speaker diarization errors	251

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		Play-By-Post Corpus	
Number of Shows	555	Number of campaigns	896
Number of Episodes	38,713	Average players per campaign	8
Number of Episodes (transcribed)	32,602	Average turns per campaign	910
Average no. of speakers per ep.	2.23	Average words per campaign	64,941
Average no. of turns per episode	597	Total turns	815,106
Average no. of words per episode	11,374	Total words	58,187,526
Total no. of turns	19,478,136	Average dice rolls per campaign	594
Total no. of words	370,824,073	Average annotations per campaign	94
		Total dice rolls	532,270
		Total annotations	84,447

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

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

(2.2 speakers per episode is certainly an underestimate given that most RPGs have a DM and 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.<sup>1</sup> 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

<sup>1</sup><https://www.dndbeyond.com/forums/d-d-beyond-general/play-by-post>

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 rules to extract state information relating to charac-






Sep 23, 2019		#51
<b>Kerrec</b> ▾		<p>Connak will immediately give warning, pointing out the wolves. He'll fire an arrow at the furthest wolf since the line of sight appears clear...</p> <p>Short BowAttack: <b>24</b> Damage: <b>7</b></p> <p>... then move his entire movement back towards the party (Left 5, Up 1).</p>
Sep 24, 2019		#52
<b>AshDrunbar</b> ▾		<p>Tabesha'l felt a small tinge of surprise: why would two wolves take on a full party of two-legs ? This whole thing was wrong, as she knew the shy nature of these creatures, who often would rather run and hide than confront humanoids, especially near settlements. Could it be that the dragon already devoured or drove away all the local game... ? For certain, she'd never seen wolves of such an aggressive nature, except during exceedingly harsh winters.</p> <p>She knew she soon wouldn't have the luxury to reflect further on these subtleties, however, and so she took a steady position with both legs firmly planted on the road, as she uttered the incantation and formed the arcane symbols she would need to protect herself from the two rabid, desperate beasts. As she felt the magical barrier rise around her, she slowly but steadily backed away from the wolves, retreating behind the rest of the group and concentrating on her next spell.</p> <p><b>Thaumaturgist</b>  <b>Location:</b> Aerth and Beyond  <b>Join Date:</b> 7/28/2018  <b>Posts:</b> 479</p> <p><b>** Casting <i>Mage Armor</i> on herself as an Action and moving 30 feet back up the road. **</b></p>
Sep 24, 2019		#53
<b>Kromp</b> ▾		<p>As Barric hears Connak's warning, he knocks an arrow to his bow, the manticores pushed to the back of his mind for now. Seeing Tabesha'l retreat behind, he stands his position in front of her, watching Vaxis getting his shield up and Brutus ready to move into the fray. He can barely see the shapes moving behind the brush, and holds his aim steady, waiting for a clearer shot.</p> <p>No movement. Readying a longbow attack when a wolf is not behind the cover of the trees. If triggered, <b>22</b> Damage: <b>11</b></p>

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

ter 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<sup>2</sup>, 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  $f(\text{current utterance, previous state, history}) \rightarrow \text{new state}$ , and response generation is a language generation task where  $f(\text{current state, history}) \rightarrow \text{next utterance}$ . 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

<sup>2</sup><https://forums.giantitp.com/forumdisplay.php?3-Play-by-Post-Games>

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  $n$ th 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<sub>1</sub> SEP TURN<sub>2</sub> SEP ... SEP TURN<sub>N-1</sub> → TURN<sub>N</sub>

### 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

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

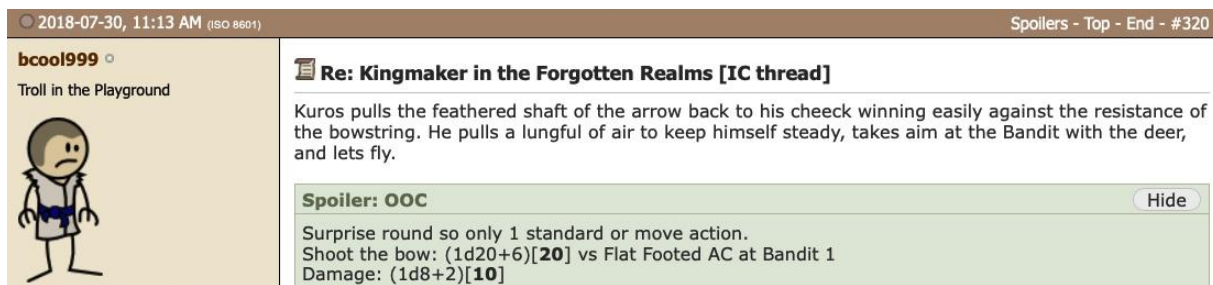


Figure 2: An example post from the Giant in the Playground forums where the text is segmented into in-character (IC) and out-of-character (OOC) portions.

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 <sup>3</sup>	Character classes perform different actions (e.g. wizards cast spells, thieves pick locks)
Character Race	D&D fantasy races <sup>4</sup>	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<sub>1</sub> SEP TURN<sub>2</sub> SEP ... SEP TURN<sub>N-1</sub> → TURN<sub>N</sub>

where TURN<sub>i</sub> denotes text along with player id for i<sup>th</sup> dialog turn and SEP is the separator. Note that the last turn TURN<sub>N</sub> 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<sub>1</sub> STATE<sub>1</sub> SEP TURN<sub>2</sub> STATE<sub>2</sub> SEP ... SEP TURN<sub>N-1</sub> STATE<sub>N-1</sub> SEP STATE<sub>N</sub> → TURN<sub>N</sub>

where STATE<sub>i</sub> denotes the set of control features for i<sup>th</sup> 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<sub>1</sub> STATE<sub>1</sub> SEP TURN<sub>2</sub> STATE<sub>2</sub> SEP ... SEP TURN<sub>N-1</sub> STATE<sub>N-1</sub> → TURN<sub>N</sub>

**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.

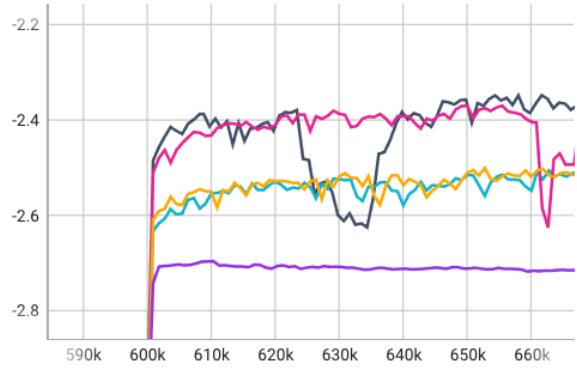


Figure 3: Perplexity of our models after pretraining on generic dialogue data for 600k steps, and then finetuning to our data for a further 60k steps

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	<b>2.34</b>	<b>49.67</b>
LLM-DND-ALL-CTRL	2.37	49.02
LLM-DND-Podcast	2.53	46.64
LLM-Podcast	2.69	44.6

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

TASK:

TURN<sub>1</sub> STATE<sub>1</sub> SEP TURN<sub>2</sub> STATE<sub>2</sub> SEP ... SEP TURN<sub>N-1</sub> STATE<sub>N-1</sub> → TURN<sub>N</sub>

**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 perplexity 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) 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 dialogue 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<sub>1</sub> STATE<sub>1</sub> SEP TURN<sub>2</sub> STATE<sub>2</sub> SEP ... SEP TURN<sub>N-1</sub> STATE<sub>N-1</sub> TURN)<sub>N</sub> → STATE<sub>N</sub>

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 prepreserved 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""	<b>Response 1</b> <b>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."</b>	Does this response make sense? <input type="checkbox"/> Makes sense <input type="checkbox"/> Does not make sense Is this response specific to the given context? <input type="checkbox"/> Specific <input type="checkbox"/> Not specific  How interesting is this response on a scale of 10? _____
	<b>Response 2</b> <b>del takes the crowbar and tells xavier to find hes ****en axe and ill keep good care of him</b>	Does this response make sense? <input type="checkbox"/> Makes sense <input type="checkbox"/> Does not make sense Is this response specific to the given context? <input type="checkbox"/> Specific <input type="checkbox"/> Not specific  How interesting is this response on a scale of 10? _____
	<b>Response 3</b> <b>Anyone see my axe, and leather?</b>	Does this response make sense? <input type="checkbox"/> Makes sense <input type="checkbox"/> Does not make sense Is this response specific to the given context? <input type="checkbox"/> Specific <input type="checkbox"/> Not specific  How interesting is this response on a scale of 10? _____
	<b>Response 4</b> <b>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."</b>	Does this response make sense? <input type="checkbox"/> Makes sense <input type="checkbox"/> Does not make sense Is this response specific to the given context? <input type="checkbox"/> Specific <input type="checkbox"/> 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 models' 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 another’s

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 Clangeddin!" 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.

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<sup>5</sup>, 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. 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

<sup>5</sup><https://avrae.io>



660 generation into two steps: ideation of the story plot,  
661 followed by the realization of sentences. This *con-*  
662 *trollable story generation* is where we are currently  
663 seeing a lot of work within neural automated story  
664 generation since it enables the neural network(s) to  
665 focus on a single task at a time.

666 Due to the conversational nature of D&D , we  
667 decided to use a dialog-based system. Deep neural  
668 networks have been used for dialog agents for a  
669 while (Serban et al., 2016), with a shift toward us-  
670 ing transformers in recent years (Zhang et al., 2019;  
671 Ghazarian et al., 2021). Like in automated story  
672 generation and other neural text generation tasks,  
673 we are also seeing controllability being an impor-  
674 tant factor being integrated into systems. This  
675 includes using deep reinforcement learning tech-  
676 niques to guide the dialog toward a goal (Li et al.,  
677 2016; Saleh et al., 2020) or controlling for style  
678 (Zhang et al., 2018; Smith et al., 2020).

679 In this paper, we use LaMDA, a transformer-  
680 based open-domain dialogue system that building  
681 on the Meena model (Adiwardana et al., 2020). The  
682 original Meena model was an end-to-end model  
683 trained on public conversations found on social me-  
684 dia. Controllable text generation with transformers  
685 has been seen before with CTRL (Keskar et al.,  
686 2019), a language model that is conditioned on a  
687 given "control code" in addition to the Bayesian  
688 history. This work takes a similar approach. We  
689 integrate contextual information such as character  
690 descriptions, actions, and in- and out-of-character  
691 classifications.

692 We have finetuned our LaMDA models on data  
693 crawled from D&D Beyond<sup>6</sup>. This data contains  
694 both in-character and out-of-character dialog and  
695 can be used in conjunction with Rameshkumar  
696 and Bailey (2020)'s dataset from Critical Role  
697 (a D&D podcast), Louis and Sutton (2018)'s  
698 dataset from roleplayerguild.com (a D&D fo-  
699 rum), and/or Urbanek et al. (2019)'s crowdsourced  
700 LIGHT dataset. For the purposes of this work,  
701 we will only be working with our D&D Beyond  
702 dataset.

## 703 9 Discussion and Conclusions

704 We find that training on D&D data results in much  
705 higher quality outputs than a vanilla dialogue sys-  
706 tem (as expected), that controlling the model to  
707 generate in-character responses results in substan-  
708 tially more interesting output, and that conditioning

<sup>6</sup><https://www.dndbeyond.com/>

709 on game state information qualitatively results in  
710 responses that are appropriate to the character class.  
711 Our preliminary experiments with using the large  
712 language models to perform game state tracking  
713 show low performance even after fine-tuning, sug-  
714 gesting that other models may be required for an  
715 AI to play D&D track the full state of the game.

## 716 10 Acknowledgments

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

## 720 References

- 721 Daniel Adiwardana, Minh-Thang Luong, David R So,  
722 Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang,  
723 Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu  
724 Quoc, and V Le. 2020. *Towards a Human-like Open-*  
725 *Domain Chatbot*. *arXiv preprint arXiv:2001.09977*.
- 726 Prithviraj Ammanabrolu, William Broniec, Alex  
727 Mueller, Jeremy Paul, and Mark O. Riedl. 2020a.  
728 *Toward Automated Quest Generation in Text-*  
729 *Adventure Games*. In *International Conference on*  
730 *Computational Creativity (ICCC)*.
- 731 Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li,  
732 Arthur Szlam, Tim Rocktäschel, and Jason Weston.  
733 2020b. *How to Motivate Your Dragon: Teaching*  
734 *Goal-Driven Agents to Speak and Act in Fantasy*  
735 *Worlds*. In *The North American Chapter of the*  
736 *Association for Computational Linguistics (NAACL-*  
737 *HLT)*.
- 738 Boyd Branch, Piotr Mirowski, and Kory Mathewson.  
739 2021. *Collaborative Storytelling with Human Ac-*  
740 *tors and AI Narrators*. In *International Conference*  
741 *on Computational Creativity (ICCC)*, pages 97–101.
- 742 Sahith Dambekodi, Spencer Frazier, Prithviraj Am-  
743 manabrolu, and Mark O Riedl. 2020. *Playing Text-*  
744 *Based Games with Common Sense*. In *Wordplay:*  
745 *When Language Meets Games Workshop (NeurIPS*  
746 *2020)*.
- 747 Simon Ellis and James Hendler. 2017. *Comput-*  
748 *ers Play Chess, Computers Play Go...Humans Play*  
749 *Dungeons & Dragons*. *IEEE Intelligent Systems*,  
750 32(4):31–34.
- 751 Angela Fan, Mike Lewis, and Yann Dauphin. 2018. *Hi-*  
752 *erarchical Neural Story Generation*. In *Proceedings*  
753 *of the 56th Annual Meeting of the Association for*  
754 *Computational Linguistics*, pages 889–898.
- 755 Angela Fan, Jack Urbanek, Pratik Ringshia, Emily  
756 Dinan, Emma Qian, Siddharth Karamcheti, Shri-  
757 mai Prabhumoye, Douwe Kiela, Tim Rocktäschel,  
758 Arthur Szlam, and Jason Weston. 2020. *Generating*  
759 *Interactive Worlds with Text*. In *AAAI Conference*  
760 *on Artificial Intelligence (AAAI)*, pages 1693–1700.

761	David Ferrucci, Eric Brown, Jennifer Chu-Carroll,	Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang,	818
762	James Fan, David Gondek, Aditya A Kalyanpur,	William Hancock, Shruti Singh, Brent Harrison, and	819
763	Adam Lally, J William Murdock, Eric Nyberg, John	Mark O. Riedl. 2018a. <a href="#">Event Representations for</a>	820
764	Prager, et al. 2010. Building watson: An overview	<a href="#">Automated Story Generation with Deep Neural Nets.</a>	821
765	of the deepqa project. <i>AI magazine</i> , 31(3):59–79.	In <i>Thirty-Second AAAI Conference on Artificial In-</i>	822
		<i>telligence</i> , pages 868–875, New Orleans, Louisiana.	823
766	Matt Gardner, Joel Grus, Mark Neumann, Oyvind	Lara J. Martin, Srijan Sood, and Mark O. Riedl. 2018b.	824
767	Taffjord, Pradeep Dasigi, Nelson Liu, Matthew Pe-	<a href="#">Dungeons and DQNs: Toward Reinforcement Learn-</a>	825
768	ters, Michael Schmitz, and Luke Zettlemoyer. 2018.	<a href="#">ing Agents that Play Tabletop Roleplaying Games.</a>	826
769	AllenNLP: A deep semantic natural language pro-	In <i>Joint Workshop on Intelligent Narrative Technolo-</i>	827
770	cessing platform. <i>arXiv preprint arXiv:1803.07640</i> .	<i>gies and Workshop on Intelligent Cinematography</i>	828
771	Sarik Ghazarian, Zixi Liu, Tuhin Chakrabarty, Xuezhe	<i>and Editing (INT-WICED)</i> , Edmonton, AB, Canada.	829
772	Ma, Aram Galstyan, and Nanyun Peng. 2021. <a href="#">DiS-</a>		
773	<a href="#">CoL: Toward Engaging Dialogue Systems through</a>		
774	<a href="#">Conversational Line Guided Response Generation.</a>	Kory W. Mathewson, Pablo Samuel Castro, Colin	830
775	<i>arXiv preprint arXiv:2102.02191</i> .	Cherry, George Foster, and Marc G. Bellemare.	831
776	Matan Haroush, Tom Zahavy, Daniel J. Mankowitz,	2019. <a href="#">Shaping the Narrative Arc: An Information-</a>	832
777	and Shie Mannor. 2018. <a href="#">Learning How Not to Act</a>	<a href="#">Theoretic Approach to Collaborative Dialogue.</a>	833
778	<a href="#">in Text-Based Games.</a> In <i>Workshop Track at ICLR</i>	<i>arXiv preprint arXiv:1901.11528</i> , page 20.	834
779	<i>2018</i> , pages 1–4.	Johannes E. M. Mosig, Vladimir Vlasov, and Alan	835
780	Chenglong Hou, Chensong Zhou, Kun Zhou, Jinan	Nichol. 2020. <a href="#">Where is the context? - A critique</a>	836
781	Sun, and Sisi Xuanyuan. 2019. <a href="#">A survey of deep</a>	<a href="#">of recent dialogue datasets.</a> <i>CoRR</i> , abs/2004.10473.	837
782	<a href="#">learning applied to story generation.</a> In <i>Smart Com-</i>		
783	<a href="#">puting and Communication</a> , pages 1–10. Springer	Xiangyu Peng, Siyan Li, Sarah Wiegrefe, and Mark	838
784	International Publishing.	Riedl. 2021. <a href="#">Inferring the Reader: Guiding Auto-</a>	839
785	Jeremy Howard and Sebastian Ruder. 2018. <a href="#">Universal</a>	<a href="#">mated Story Generation with Commonsense Reason-</a>	840
786	<a href="#">language model fine-tuning for text classification.</a> In	<i>Workshop on Narrative Understanding at</i>	841
787	<i>Proceedings of the 56th Annual Meeting of the As-</i>	<i>NAACL-HLT 2021</i> .	842
788	<i>sociation for Computational Linguistics (Volume 1:</i>	Shrimai Prabhumoye, Khyathi Raghavi Chandu, Rus-	843
789	<i>Long Papers)</i> , pages 328–339, Melbourne, Australia.	lan Salakhutdinov, and Alan W Black. 2019. <a href="#">"My</a>	844
790	Association for Computational Linguistics.	<a href="#">Way of Telling a Story": Persona based Grounded</a>	845
791	Daphne Ippolito, David Grangier, Chris Callison-	<a href="#">Story Generation.</a> In <i>Storytelling Workshop at ACL</i>	846
792	Burch, and Douglas Eck. 2019. <a href="#">Unsupervised Hier-</a>	<i>2019</i> , pages 11–21.	847
793	<a href="#">archical Story Infilling.</a> In <i>Proceedings of the First</i>	Revanth Rameshkumar and Peter Bailey. 2020. <a href="#">Story-</a>	848
794	<i>Workshop on Narrative Understanding</i> , pages 37–	<a href="#">telling with Dialogue: A Critical Role Dungeons and</a>	849
795	43.	<a href="#">Dragons Dataset.</a> In <i>Annual Meeting of the Asso-</i>	850
796	Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino,	<i>ciation for Computational Linguistics (ACL)</i> , pages	851
797	Richard Socher, and Hal Daumé III. 2014. A neural	5121–5134.	852
798	network for factoid question answering over para-	Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and	853
799	graphs. In <i>Proceedings of the 2014 conference on</i>	Jianfeng Gao. 2020. <a href="#">PlotMachines: Outline-</a>	854
800	<i>empirical methods in natural language processing</i>	<a href="#">Conditioned Generation with Dynamic Plot State</a>	855
801	<i>(EMNLP)</i> , pages 633–644.	<a href="#">Tracking.</a> In <i>Conference on Empirical Methods</i>	856
802	Nitish Shirish Keskar, Bryan McCann, Lav R. Varsh-	<i>in Natural Language Processing (EMNLP)</i> , pages	857
803	ney, Caiming Xiong, and Richard Socher. 2019.	4274–4295.	858
804	<a href="#">CTRL: A Conditional Transformer Language Model</a>	Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin	859
805	<a href="#">for Controllable Generation.</a> <i>arXiv preprint</i>	Sivaraman, Rezvaneh Rezapour, and Rosie Jones.	860
806	<i>arXiv:1909.05858</i> .	2021. <a href="#">Detecting Extraneous Content in Podcasts.</a>	861
807	Jiwei Li, Will Monroe, Alan Ritter, and Dan Jurafsky.	In <i>European Chapter of the Association for Compu-</i>	862
808	2016. <a href="#">Deep Reinforcement Learning for Dialogue</a>	<i>tational Linguistics (EACL)</i> , pages 1166–1173, On-	863
809	<a href="#">Generation.</a> In <i>Conference on Empirical Methods in</i>	line. Association for Computational Linguistics.	864
810	<i>Natural Language Processing</i> , pages 1192–1202.	Melissa Roemmele. 2018. <a href="#">Neural Networks for Narra-</a>	865
811	Annie Louis and Charles Sutton. 2018. <a href="#">Deep Dun-</a>	<a href="#">tive Continuation.</a> Ph.D. thesis, University of South-	866
812	<a href="#">geons and Dragons: Learning Character-Action In-</a>	thern California.	867
813	<a href="#">teractions from Role-Playing Game Transcripts.</a> In	Abdelrhman Saleh, Natasha Jaques, Asma Ghande-	868
814	<i>Conference of the North American Chapter of the</i>	harioun, Judy Hanwen Shen, and Rosalind Picard.	869
815	<i>Association for Computational Linguistics: Human</i>	2020. <a href="#">Hierarchical Reinforcement Learning for</a>	870
816	<i>Language Technologies (NAACL-HLT)</i> , pages 708–	<a href="#">Open-Domain Dialog.</a> In <i>AAAI Conference on Ar-</i>	871
817	713.	<i>tificial Intelligence (AAAI)</i> , pages 8741–8748.	872

873	Abigail See, Aneesh Pappu, Rohun Saxena, Akhila	Shunyu Yao, Rohan Rao, Matthew Hausknecht, and	931
874	Yerukola, and Christopher D. Manning. 2019. <a href="#">Do</a>	Karthik Narasimhan. 2020. <a href="#">Keep CALM and Ex-</a>	932
875	<a href="#">Massively Pretrained Language Models Make Better</a>	<a href="#">plore: Language Models for Action Generation in</a>	933
876	<a href="#">Storytellers?</a> In <i>Conference on Computational Nat-</i>	<a href="#">Text-based Games</a> . In <i>Conference on Empirical</i>	934
877	<i>ural Language Learning (CoNLL)</i> , pages 843–861.	<i>Methods in Natural Language Processing (EMNLP)</i> ,	935
		pages 8736–8754.	936
878	Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and	Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur	937
879	Joelle Pineau. 2016. <a href="#">Generative Deep Neural Net-</a>	Szlam, Douwe Kiela, and Jason Weston. 2018. <a href="#">Per-</a>	938
880	<a href="#">works for Dialogue: A Short Review</a> . In <i>Con-</i>	<a href="#">sonalizing Dialogue Agents: I have a dog, do you</a>	939
881	<i>ference on Neural Information Processing Systems</i>	<a href="#">have pets too?</a> In <i>Annual Meeting of the Association</i>	940
882	<i>(NeurIPS), Workshop on Learning Methods for Dia-</i>	<i>for Computational Linguistics (ACL)</i> , pages 2204–	941
883	<i>logue</i> , pages 1–6.	2213, Melbourne, Australia.	942
884	Eric Michael Smith, Diana Gonzalez-Rico, Emily	Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun	943
885	Dinan, and Y-Lan Boureau. 2020. <a href="#">Controlling</a>	Chen, Chris Brockett, Xiang Gao, Jianfeng Gao,	944
886	<a href="#">Style in Generated Dialogue</a> . <i>arXiv preprint</i>	Jingjing Liu, and Bill Dolan. 2019. <a href="#">DialoGPT:</a>	945
887	<i>arXiv:2009.10855</i> .	<a href="#">Large-Scale Generative Pre-training for Conver-</a>	946
		<a href="#">sational Response Generation</a> . <i>arXiv preprint</i>	947
888	Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J.	<i>arXiv:1911.00536</i> .	948
889	Martin, Animesh Mehta, Brent Harrison, and		
890	Mark O. Riedl. 2019. <a href="#">Controllable Neural Story</a>	<b>A Lost Mine of Phandelver Adventure</b>	949
891	<a href="#">Plot Generation via Reinforcement Learning</a> . In		
892	<i>International Joint Conference on Artificial Intelli-</i>	Here is an excerpt from the adventure book that	950
893	<i>gence (IJCAI)</i> , Macau, China.	the Dungeon Master was using in our example	951
894	Romal Thoppilan, Daniel De Freitas, Jamie Hall,	game play. The adventure book provides boxed	952
895	Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze	text, which is descriptive text to be read aloud to	953
896	Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du,	paraphrased. It also gives details about the com-	954
897	YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng,	bat that is about to ensure, and links to relevant	955
898	Amin Ghafouri, Marcelo Menegali, Yanping Huang,	game rules (like stealth checks, and statistics about	956
899	Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao	the monsters that the characters will be in combat	957
900	Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts,	with).	958
901	Maarten Bosma, Yanqi Zhou, Chung-Ching Chang,	The adventure begins as the player characters are	959
902	Igor Krivokon, Will Rusch, Marc Pickett, Kath-	escorting a wagon full of provisions and supplies	960
903	leen S. Meier-Hellstern, Meredith Ringel Morris,	from Neverwinter to Phandalin. The journey takes	961
904	Tulsee Doshi, Renelito Delos Santos, Toju Duke,	them south along the High Road to the Triboar	962
905	Johnny Soraker, Ben Zevenbergen, Vinodkumar	Trail, which heads east (as shown on the overland	963
906	Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen	map). When they’re a half-day’s march from Phan-	964
907	Olson, Alejandra Molina, Erin Hoffman-John, Josh	dalin, they run into trouble with goblin raiders from	965
908	Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna,	the Cragmaw tribe.	966
909	Matthew Lamm, Viktoriya Kuzmina, Joe Fenton,	Read the boxed text when you’re ready to start.	967
910	Aaron Cohen, Rachel Bernstein, Ray Kurzweil,	If you create a different adventure hook, skip to	968
911	Blaise Aguera-Arcas, Claire Cui, Marian Croak,	the second paragraph and adjust the details as nec-	969
912	Ed H. Chi, and Quoc Le. 2022. <a href="#">LaMDA: Lan-</a>	essary, ignoring the information about driving the	970
913	<a href="#">guage Models for Dialog Applications</a> . <i>CoRR</i> ,	wagon.	971
914	abs/2201.08239.		
915	Jack Urbanek, Angela Fan, Siddharth Karamcheti,	In the city of Neverwinter, a dwarf	972
916	Saachi Jain, Samuel Humeau, Emily Dinan, Tim	named Gundren Rockseeker asked you	973
917	Rocktäschel, Douwe Kiela, Arthur Szlam, and Ja-	to bring a wagonload of provisions to	974
918	son Weston. 2019. <a href="#">Learning to Speak and Act in a</a>	the rough-and-tumble settlement of Phan-	975
919	<a href="#">Fantasy Text Adventure Game</a> . In <i>Conference on</i>	dalin, a couple of days’ travel southeast	976
920	<i>Empirical Methods in Natural Language Process-</i>	of the city. Gundren was clearly excited	977
921	<i>ing and the International Joint Conference on Natu-</i>	and more than a little secretive about	978
922	<i>ral Language Processing (EMNLP-IJCNLP)</i> , pages	his reasons for the trip, saying only that	979
923	673–683, Hong Kong, China. Association for Com-	he and his brothers had found “some-	980
924	putational Linguistics.	thing big,” and that he’d pay you ten	981
925	Nick Walton. 2019. <a href="#">AI Dungeon</a> .		
926	Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin		
927	Knight, Dongyan Zhao, and Rui Yan. 2019.		
928	<a href="#">Plan-And-Write: Towards Better Automatic Story-</a>		
929	<a href="#">telling</a> . In <i>AAAI Conference on Artificial Intelli-</i>		
930	<i>gence (AAAI)</i> , volume 33, pages 7378–7385.		

982	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."		
983			
984			
985			
986			
987			
988			
989	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.		
990			
991			
992			
993			
994			
995			
996	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.		
997			
998			
999			
1000			
1001			
1002			
1003			
1004			
1005	Four goblins are hiding in the woods, two on each side of the road. They wait until someone approaches the bodies and then attack.		
1006			
1007	This will likely be the first of many combat encounters in the adventure. Here are the steps you should follow to run it effectively:		
1008			
1009			
1010			
1011	<ul style="list-style-type: none"> <li>Review the goblin stat block in appendix B. Since the goblins are hiding, you'll need to know their Stealth skill modifier: +6.</li> </ul>		
1012			
1013			
1014	<ul style="list-style-type: none"> <li>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).</li> </ul>		
1015			
1016			
1017			
1018			
1019			
1020			
1021			
1022			
1023			
1024			
1025			
1026	<ul style="list-style-type: none"> <li>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.</li> </ul>		
1027			
1028			
1029			
		<ul style="list-style-type: none"> <li>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</li> </ul>	1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040
		<b>B Annotation Guidelines</b>	1041
		<b>B.1 Annotation task</b>	1042
		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.	1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053
		For each response, you would be asked the following questions.	1054 1055
		<ul style="list-style-type: none"> <li>Does the response make sense? <ul style="list-style-type: none"> <li>Use your common sense here. Is the response completely reasonable in terms of the rules of D&amp;D ?</li> <li>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.</li> <li>If anything seems off—not fluent, confusing, illogical, out of context, or wrong according to the rules of D&amp;D —then rate it as Does not make sense. If in doubt, choose Does not make sense.</li> </ul> </li> </ul>	1056 1057 1058 1059 1060 1061 1062 1063 1064 1065 1066 1067 1068 1069
		<ul style="list-style-type: none"> <li>Is the response specific? <ul style="list-style-type: none"> <li>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?</li> </ul> </li> </ul>	1070 1071 1072 1073 1074 1075



1076 – The response is "specific" if it flows log- 1122  
 1077 ically from the narrative established by 1123  
 1078 the prior context. 1124  
 1079 \* Note: It is possible for a response to 1125  
 1080 "make sense" (due to being cohesive, 1126  
 1081 consistent and plausible in and of it- 1127  
 1082 self), but be marked "not specific" 1128  
 1083 when it is not a logical next step in 1129  
 1084 the overall game progression. 1130  
 1085 \* Note: "Specific" for the purposes of 1131  
 1086 this task does not have to do with 1132  
 1087 how detailed the response is per se;  
 1088 a response can be fairly general in  
 1089 its language, but still qualify as "spe-  
 1090 cific" when it is a logical next step in  
 1091 the overall game progression.

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 expo-  
 sure to D&D through TV or other channels. All 6  
 answered that they were fans of the fantasy genre.

- 1092 ● How interesting is the response?
- 1093 – You may be asked to score the response
- 1094 for its interestingness on a scale of 10.
- 1095 Choose a high score for “Interesting” if
- 1096 the response would likely catch some-
- 1097 one’s attention or arouse curiosity in the
- 1098 game; or it is insightful, creative, or
- 1099 witty with respect to the game. If the
- 1100 response is monotonous and predictable,
- 1101 or if you’re unsure, then it is Less Inter-
- 1102 esting.

**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:

- 1108 1. Have you ever played Dungeons and Dragons
- 1109 or another role playing game before?
- 1110 2. If so,
- 1111 ● roughly how many times have you
- 1112 played
- 1113 ● were you a player or a game master or
- 1114 both
- 1115 3. If not,
- 1116 ● what kind of exposure do you have to
- 1117 Dungeons and Dragons? (For example,
- 1118 have you seen it referred to in TV or
- 1119 movies)
- 1120 4. Are you a fan of the fantasy genre (like Lord
- 1121 of the Rings)?