Lore Coherent Encounter Generation for Dungeons and Dragons: LLM Fine-tuning and Benchmark

Aravinth Sivaganeshan, Nisansa de Silva, Akila Peiris

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Lore Coherent Encounter Generation for Dungeons and Dragons: LLM Fine-tuning and Benchmark

Aravinth Sivaganeshan, Nisansa de Silva, Akila Peiris
Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{sivaganeshan.22, NisansaDdS, akila.21}@cse.mrt.ac.lk

Abstract

Swathes of tasks that were erstwhile handled by other deep learning models are being taken over by Large Language Models (LLMs). While they may demonstrate reasonable results in the zero-shot configuration for most domains, in the contexts of more niche or esoteric domains, instruction tuning them on the domain at hand has shown to be effective and sometimes necessary. Dungeons & Dragons (D&D) is the most commercially successful Tabletop Role Playing Game (TTRPG) with its own unique lore mostly set in the fantasy domain. The players are confronted with fantastical *monsters* in mathematically balanced *encounters* overcoming which, contribute to the calculation of player progress. Even with the plethora of information available for D&D (or perhaps rather due to its abundance), the generation of an encounter that is coherent with the lore is a time-consuming and difficult process as there are no support tools available for the selection of coherent monsters. Recognizing this gap, we instruction-tuned a Mistral-based LLM that can function as an assistant on this matter, using instructions generated from publicly available D&D datasets. Next, we conducted a number of prompt engineering experiments on the trained LLM, such that the output from the LLM would be a list of coherent monsters when a candidate monster is given in the input. The generated outputs were examined for the coherency of lore, theme, and environment. It was observed that the outputs were partially or fully coherent with the lore in about 66.0% of the 241 candidate monsters tested.

1 Introduction

Language models trained on very large datasets are shown to have high capabilities (Brown et al., 2020; Chowdhery et al., 2022). However, these models are trained on datasets generated by humans with various goals (Rafailov et al., 2023). Therefore, the performance of LLMs might not be desirable for specific downstream tasks or specific domains. While pre-training LLMs entirely on a specific domain data is costly and resource intensive (Cottier et al., 2024), techniques such as instruction fine-tuning are quite useful in this regard.

Dungeons and Dragons (D&D) is a very popular open-ended, Table-Top, Role-Playing Game (TTRPG). It is commercially available since 1974 (Gygax and Arneson, 1974) and it is currently in its 5th edition (Crawford et al., 2014b). D&D has a set of predefined rules and there are several settings in D&D (Peiris and de Silva, 2022, 2023; Squire, 2007; Weerasundara and de Silva, 2024). Each setting has lore that describes the historical and current status of the game world[1].

In a D&D game, combat encounter is one of the most important component, through which players attain progress (Crawford et al., 2014a). In a combat encounter, players are pitted against domain specific entities called *monsters*. These monsters may be considered as either *bosses* or *minions* and are partially defined by their numerical statistics as shown in Figure 1 comparing two boss monsters (Mind Flayer[2] and Red Dragon[3]) and two minions (Intellect Devourer[4] and Kobold[5]).

The generated encounters need to align with the lore to preserve immersion and verisimilitude (Stern, 2002). As a generic example, a party going through a forest being attacked by a lion is an encounter that aligns with the lore. On the

[1]The word *World* is used as an encompassing term that may mean anything from a small region of land (Perkins et al., 2015) to a planet (Crawford et al., 2019) to a universe (Lee et al., 2019) or a multiverse (Arman et al., 2023) depending on the specific lore.

[2]https://www.dndbeyond.com/monsters/5195125-mind-flayer

[3]https://www.dndbeyond.com/monsters/5194875-adult-red-dragon

[4]https://www.dndbeyond.com/monsters/5195088-intellect-devourer

[5]https://www.dndbeyond.com/monsters/16939-kobold

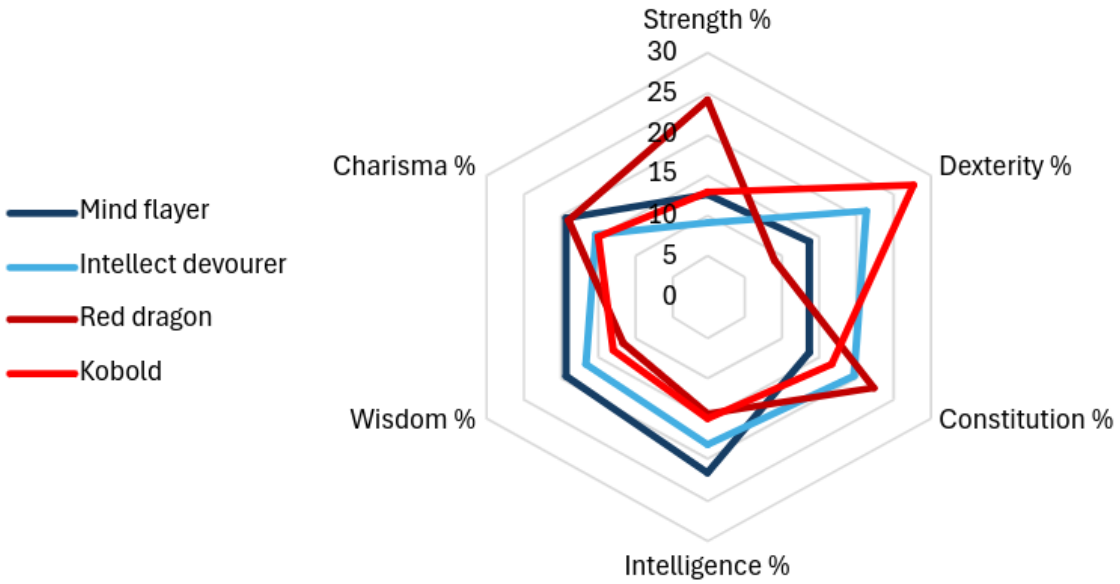

Figure 1: Comparison of creature statistics for two boss monsters (Mind Flayer and Red Dragon) and two minions (Intellect Devourer and Kobold)

other hand, a party going in a boat being attacked by a lion is an encounter that does not align with the lore. In D&D, one player enacts a special role named the *Dungeon Master* (DM) who takes the responsibility of being the lead story teller. Thus, the time-consuming task of selecting monsters for encounters according to the desired theme and difficulty is one of the responsibilities of the DM.

The *Challenge Rating* (CR) of a monster is an estimation of the threat posed by the particular monster and there are a few tools already available for calculating the difficulty of the encounter based on the challenge rating of the monsters to a particular party of players. These tools can help DM with the mathematical aspect of an encounter, but offers no help in the consistency of lore. In fact, there are no existing tools for the automatic generation of encounters with the consideration of the written lore. Currently, the DMs are expected to have near encyclopedic knowledge about the lore by themselves to make sure that aspect of the encounter is sound. Therefore, there is an existing necessity for a tool that can be used as a Dungeon master's assistant to select monsters according to the environment or to select monsters that are coherent with each other as described in the lore. Also, we analyzed the necessity by conducting a survey on reddit whether a tool with this functionality is preferable to DMs. Figure 2 shows the results of the survey. However, we had to end the poll early due to the AI related rules on the particular subreddit[6]. Therefore the poll at the time of closure had a total of 40 responses. According to the best of our knowledge, this is the first work on the specific problem. This work mainly focuses on generating coherent encounters consistent with the lore for 5th edition of D&D (5e). The language models that are currently used for the general domain cannot be used for a fantasy domain such as D&D given that a significant portion of jargon does not make sense in the general domain. Even when they do, the semantics of the words may be quite different (Peiris and de Silva, 2022). As a solution for this, we propose converting the problem of the abundance of the D&D lore into the solution itself by instruction tuning LLMs using automatically generated questions and answers from the lore documents.

For this study, we selected `Mistral7BInstruct v0.2` which is made by instruct tuning `Mistral 7B` (Jiang et al., 2023) as the base model and instruct tuned it with a set of domain specific datasets that we generated to obtain a set of models. After that, prompting experiments were done on all these models to select the best model and the best prompt. Then, we selected the final model with the best prompting technique to list the encounter and tested the best model extensively with 241 prompts. The LLM outputs from the best model for these 241 prompts were judged by 2 humans and 3 LLMs. All the links to our instruction-tuned models and

[6] https://www.reddit.com/mod/DnD/rules/

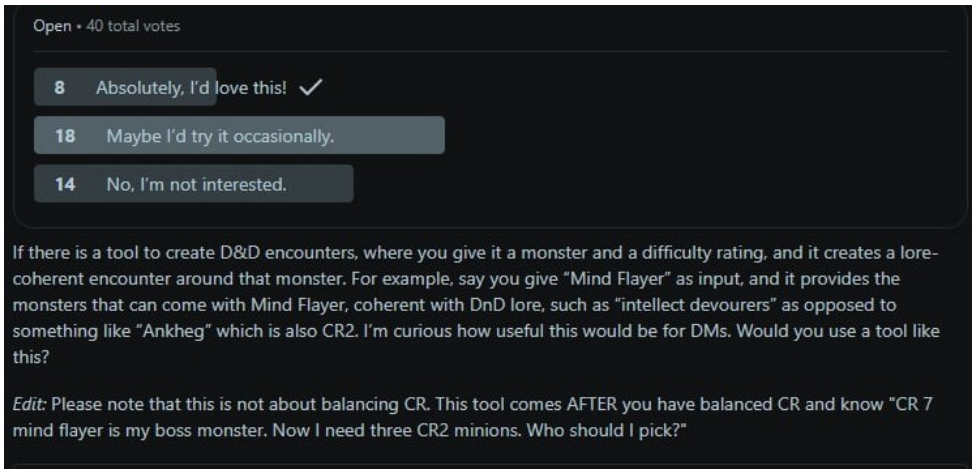

Figure 2: Reddit Survey taken for identifying whether the DMs would prefer a tool to generate encounters

the data used for instruction-tuning are given in Section 3.

2 Related Works

2.1 Existing Data Sets and Research on D&D

There is a considerable amount of official lore about D&D spread among copyrighted and non-copyrighted publications. The largest free and publicly accessible repository of relevant lore can be found on the Forgotten Realms Wiki[7] which is a Wikipedia-style collection of the lore written in the in-world perspective. For example, the article on *Mordenkainen* in ForgottenRealms wikia[8] starts as "Mordenkainen was a prolific archmage..." as opposed to the article[9] on Wikipedia for the same character which starts with "Mordenkainen is a fictional wizard...". Using the data from the wikia, Peiris and de Silva (2022) created the `FRW-dataset`[10] collection that includes 11 datasets at various types and levels on preprecessing applied to the data. This collection includes `FRW-alpaca.jsonl` which is a dataset of 41106 instructions and outputs in the alpaca format (Taori et al., 2023), which can be directly used for instruction tuning. This dataset contains instructions asking for descriptions as well as specific questions on events, places, monsters, and other concepts in D&D.

Monster statistics, which is not completely available on the *ForgottenRealms* wikia, can be accessed from `5etools`[11]. A full data dump of around 2079 monsters with their characteristics tabulated as a 32-column table can be downloaded as a csv file from `5etools`. This data can be used for generating instructions related to the mechanical aspect of the game.

Finally, the official source, `DnDbeyond`[12] contains a large amount of data. However, they do not provide direct access to any of the data (including the non-copyrighted data) in any form other than viewing on a web browser. Therefore, this source can only be used as a reference for our human experts to learn about the game and not as a data source.

There have been a number of studies using D&D data, especially in dialogue and discourse analysis (Rameshkumar and Bailey, 2020; Callison-Burch et al., 2022; Louis and Sutton, 2018). Further works have also explored the possibility of AI playing the game as DMs or players (Ellis and Hendler, 2017; Martin et al., 2018). There are some studies (Weerasundara and de Silva, 2023; Sivaganeshan and De Silva, 2023) that focus on the extended Named Entity Recognition (NER) task of identifying Dungeons and Dragons entities from text. Image (Weerasundara and de Silva, 2024) and adventure (Peiris and de Silva, 2022) generation are also aspects that have been explored for D&D. But, up to now, there are no existing works regarding the automated generation of lore consistent encounters for Dungeons and Dragons. The only available

[7] https://forgottenrealms.fandom.com
[8] https://forgottenrealms.fandom.com/wiki/Mordenkainen
[9] https://en.wikipedia.org/wiki/Mordenkainen
[10] https://huggingface.co/datasets/Akila/ForgottenRealmsWikiDataset
[11] https://5e.tools/
[12] https://www.dndbeyond.com/

tools [13] [14] [15] for encounter building only considers the mathematical aspect of the encounters.

2.2 Low Rank Adaptation

Low Rank Adaptation (LoRA) (Hu et al., 2021) is useful for fine-tuning large models to downstream tasks without updating all the parameters of the existing model. Also, this method creates a separate `adapter` for the base model for the particular application. This is useful for reducing the cost of fine-tuning and also different adapters can be made for different tasks for a single base model. Further, an extended method named `QLoRa` (Dettmers et al., 2024) can be used to reduce the memory requirements of fine-tuning by quantizing the pretrained model to 4 bits and adding a small set of LoRA weights that are tuned by backpropagating gradients through the quantized weights.

2.3 Instruction tuning LLMs for domain specific downstream tasks

Instruction tuning is a computationally effective process for adapting an LLM to a specific domain without extensive retraining or architectural changes (Zhang et al., 2023). In this technique, the LLMs are further trained using `(INSTRUCTION, OUTPUT)` pairs such that `INSTRUCTION` denotes human instruction to the model and `OUTPUT` denotes the expected output.

LoRA-based methods can be used to further improve the computational efficiency of instruction tuning. `Alpaca` (Taori et al., 2023) dataset format is a standard structure for instruction tuning which represents the instructions as a `JSON` array with objects containing `(INSTRUCTION,INPUT,OUTPUT)`. There are several tools and frameworks for this purpose where `Axolotl`[16] is one of easy-to-use tool.

2.4 Prompt Engineering

The *prompt* is the input provided to the LLM to obtain the output. Empirically, it is shown that better prompts lead to better outputs across different tasks (Wei et al., 2022; Liu et al., 2023). Due to the growth and widespread use of LLMs, prompting has become an emerging field. Text-based prompting can be generally divided into categories of: In-Context Learning, thought generation, decomposition, ensembling, and self-criticism (Schulhoff et al., 2024).

2.5 LLM-as-a-Judge

LLMs are a compelling alternative to traditional expert driven evaluations due to their ability to process diverse data types and provide scalable, flexible and consistent assessments (Gu et al., 2024). There are multiple works where LLMs replace human judges or used together with human judges for rapid, scalable evaluation (Ashktorab et al., 2024; Bavaresco et al., 2025; Tseng et al., 2024).

3 Methodology

This section provides an overview of the collection and preparation of instruction datasets, instruction tuning `MistralInstructv0.2`[17] with the instruction dataset, prompt engineering and evaluation of fine-tuned LLM outputs.

3.1 Building the Instruction Datasets

A proper instruction dataset is essential for adapting an LLM to the user objective and the D&D domain. We collected and processed data from numerous publicly available sources. First, 41106 instructions were obtained from `FRW-J-Alpaca.jsonl` discussed in section 2.1. Let us call this instruction data set `FRW-I`.

5et-I Instruction Dataset: As further discussed in section 2.1, in order to obtain information that is not included in Forgotten Realms Wiki (and thus not in `FRW-I`), we use the data export from `5etools`. This includes the data on feature columns such as environment, size, alignment, type, speed, strength, and also contains descriptions of several features such as traits, actions, bonus actions and lair actions. Given that these are also traits that are intrinsic to the given monsters in the D&D domain, these features may impact the decision of whether a set of monsters can come together in an encounter.

Linguistic diversity of instructions is seen to help models generalize better (Zhang et al., 2024). So, we formed questions in 3-5 linguistically different formats for each column. Next, when building an instruction entry, for each monster we picked a random format out of the different formats of questions we created. Thus, an instruction in the alpaca

[13] https://www.dndbeyond.com/encounter-builder
[14] https://www.aidedd.org/dnd-encounter/
[15] https://www.kassoon.com/dnd/5e/generate-encounter/
[16] https://github.com/axolotl-ai-cloud/axolotl
[17] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

```
1. Describe the regional effects of {}?
2. Explain the effects that {} can have in its region?
3. Describe the effects that {} can have in its surroundings.
4. Explain the impact that {} can have in its surroundings.
5. Describe the impact that {} can have in its surroundings.
```

Prompt Template 1: Example Question Formats for Regional Effects of a Monster

```
{
   "instruction": "What are the monsters that can be together
        with an Ancient Red Dragon in an encounter?",
   "input": "",
   "output": "Kobolds, Dragon Cultists, Red Dragon Wyrmlings"
}
```

Prompt Template 2: Example of an instruction built from an encounter

format was then generated with the question as the instruction, with empty input and the column value as the output. In a similar way instructions were created from questions such that the environment is in the instruction and all relevant monsters in the output. In total, this process yielded an instruction tuning dataset with a total of 27,959 instructions using the data on 21 columns. Let us call this instruction data set 5et-I.

5et-I-All Instruction Dataset: For comparison in fine-tuning, instead of selecting a random question format when building an instruction entry, we used all the question formats created for a column to each column and a monster pair. From this, 3-5 instructions were formed for a column, monster pair. This process yielded a total of 110,089 instructions. Let us call this instruction data set `5et-I-All`.

Enc-I Instruction Dataset: Associations between monsters is a main factor that impacts in deciding whether a set of monsters can come together in an encounter. Thus, mining association rules from human expert-created encounters and making use of them in building an instruction set could be considered as useful.

For this, a total of 3786 human expert-created encounters were extracted from publicly available data from numerous online sources. From this, a total of 2587 monster itemsets that can come together in an encounter were obtained.

Then, frequent itemset mining was done using `Apriori` algorithm on the above 2587 itemsets to extract association rules from the encounter itemsets. Starting from higher support and confidence, association rules were obtained, manually examined and then support and confidence were reduced to get more association rules. Finally, 1096 association rules were obtained when the support is 0.0003 and confidence is 0.2. In this process, support was even tested with the very low value of 0.0003. The reason for this is because the itemset list is not from a balanced list of encounters, given that we sourced it from only the publicly available encounters. Thus, it is not meaningful to reject encounters looking for a higher support threshold. On the other hand, considering the domain, threshold of 0.2 was found to be acceptable. With this process, a list of 1096 association rules was finalized. Building the instructions from this was reasonably similar to the process used earlier to build `5et-I` using different question templates and selecting a random one. Prompt Template 2 shows an example of an instruction built from an association rule. Let us call this instruction data set `Enc-I`.

Enc-I-All Instruction Dataset: A similar process to creating `Enc-I` was followed with the only difference of using all the question formats created instead of using a random format. From this process, an instruction dataset of 3288 instructions was created. Let us call this data set `Enc-I-All`.

Aggregated Instruction Datasets: The instruction datasets `FRW-I`, `5et-I`, and `Enc-I` were merged into a single file named `FRW-dnd-encounter dataset`[18] containing 70161 instructions in alpaca format and the instruction order was randomized to make the instruction dataset more suitable for training. Similarly, instruction datasets `FRW-I`, `5et-I-All` and `Enc-I-All` were merged into a single file named `FRW-dnd-encounter-all dataset`[19] with 154,483 instructions and the instruction order was randomized. Further, based on the results after trying different prompts, it was found that role prompting worked better in comparison to the other prompt formats. Based on that, another dataset was created from `FRW-dnd-encounter dataset`. Every instruction in the dataset was prefaced with

[18] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools.jsonl

[19] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools_all.jsonl

```
{
  "instruction": "You are a D&D expert. Provide an answer to
       the following question.",
  "input": "To which types of damage is an Aboleth Spawn
       resistant?",
  "output": "psychic damage"
}
```

Prompt Template 3: Example of an instruction in the FRW-dnd-encounter-role dataset

"*You are a D&D expert. Provide an answer to the following question*". The questions which were previously added as "*instruction*" were now changed to "*input*". By doing this, a new dataset was created named `FRW-dnd-encounter-role dataset`[20]. Prompt Template 3 shows an example of an instruction in this dataset. Similarly, the `FRW-dnd-encounter-role-all dataset`[21] is created by modifying `FRW-dnd-encounter-all dataset`.

3.2 Training the model

`Mistral7BInstructv0.2`[22] was taken as the base model to be fine-tuned with the D&D instruction dataset. This is due to the fact that the Mistral-Instruct model (Jiang et al., 2023) with 7 billion parameters, has been shown to be efficient and having comparable performance to the state of the art 13B parameter chat models (Jiang et al., 2023). Also, it has been shown to work well with `LORA` (Fujiwara et al., 2024). `Quantized LORA (QLORA)` is utilized to reduce the computing resources used for training. The training was done using `Axolotl` framework in an H100 2XM GPU cloud virtual machine from runpod.io[23] with 80 GB VRAM. 8 models were obtained by finetuning the base model, `Mistral7BInstructv0.2` with the generated instruction datasets. `Model1v0.1`[24] and `Model1v0.2`[25] were obtained by finetuning with instruction dataset `FRW-dnd-encounter` for 1 and 2 epochs respectively. Similarly, another 6 models were obtained by finetuning with the generated datasets for 1 and 2 epochs. `Model2v0.1`[26] and `Model2v0.2`[27] were obtained by finetuning with `FRW-dnd-encounter-role`. `Model3v0.1`[28] and `Model3v0.2`[29] were obtained by finetuning with the `FRW-dnd-encounter-all dataset`.`Model4v0.1`[30] and `Model4v0.2`[31] were from finetuning with the `FRW-dnd-encounter-role-all dataset`. QLoRa training configurations for fine-tuning every model available in the link given with each of the above models.

3.3 Prompt Engineering

Prompt engineering is a crucial task for optimizing the performance of a large language model on customized tasks (Schulhoff et al., 2024). The prompt types that were tried were based on role prompting, zero-shot prompting, few-shot prompting, self criticism and chain of thought reasoning. Prompt Template 4 shows the 10 prompt variations that were tested for encounter generation. These prompt engineering methods were taken from Schulhoff et al. (2024). As the first step, all the 8 finetuned models were tested with the basic prompts (Prompt 1) and (Prompt 2) for 20 frequently used monsters and the best 2 models were selected. Then, for these 2 models, 10 prompt types were tried out and tested with the task of asking the LLM to create encounters for 20 frequently used monsters that were picked from diverse categories.

The prompt and the best model that yielded the best result was selected, and the best model was tested with the best prompt format for 241 frequently used monsters D&D. Then, the model outputs given for the 241 monsters were evaluated by 2 human annotators and 3 LLMs. `GPT-4.1`, `Gemini-2.5-flash` and `DeepSeek-V3-0324` were the LLMs used as annotators. Also, different prompt structures were tried on LLMs, asking the LLM to provide judgements whether the model outputs are correct, partially correct or wrong. The prompt format for which the LLM answers show

[20] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools_role.jsonl

[21] https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/blob/main/FRW-J_and_dnd_5etools_role_all.jsonl

[22] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[23] https://www.runpod.io/

[24] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2v0.1

[25] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2v0.2

[26] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2rolev0.1

[27] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2rolev0.2

[28] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2allv0.1

[29] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2allv0.2

[30] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2roleallv0.1

[31] https://huggingface.co/Aravinth92/Mistral_v0.2_dnd5etools2roleallv0.2

Prompt Formats

1. **Prompt 1 (Zero shot basic prompt) :** Create an encounter that consists of a {}?
2. **Prompt 2 (Prompt with the same format as training instruction) :** What are the minion monsters that can come with an {} in an encounter?
3. **Prompt 3 (Prompt with the same format as training instruction :** Can you tell me the monsters that can go together with a {} in an encounter?
4. **Prompt 4 (Prompt asking for explanation) :** Give me the answer with explanation. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with.
5. **Prompt 5 (Prompt asking to walk through thinking process of LLM) :** Walk me through this context in manageable parts step by step, summarising and analysing as we go. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {}, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
6. **Prompt 6 (Repeating prompt in answer to enable generation) :** Repeat the following prompt in your answer. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
7. **Prompt 7 (Role prompting) :** You are a D&D expert. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
8. **Prompt 8 (Role prompt with repeating the prompt in answer) :** You are a D&D expert. Repeat the prompt in your answer. A D&D encounter had a Mind flayer and 3 Intellect Devourers. If I replace the Mind Flayer with a {}, ignoring the CR difference; but adapting the encounter theme to match the creature type and the typical environment of a {} lair, give me 4 minion monsters that I could use to replace the Intellect Devourers with?
9. **Prompt 9 (Role prompt with examples in prompt) :** You are a D&D expert planning to create coherent D&D encounters. A coherent D&D encounter with a Mind flayer as the boss monster has Intellect Devourers and Thralls as minions. A coherent D&D encounter with a Red Dragon as the boss monster has Kobolds and dragon cultists as minions. Give a similarly coherent D&D encounter with a {} as the boss monster while ignoring the CR difference, but adapting the encounter theme to match the creature type and the typical environment of a {} lair.
10. **Prompt 10 (Role prompt with specifing output) :** You are a D&D expert planning to create coherent D&D encounters. A coherent D&D encounter with a Mind flayer as the boss monster has Intellect Devourers and Thralls as minions. A coherent D&D encounter with a Red Dragon as the boss monster has Kobolds and dragon cultists as minions. Give a similarly coherent D&D encounter with a {} as the boss monster and 4 different minion monsters while ignoring the CR difference, but adapting the encounter theme to match the creature type and the typical environment of a {} lair.

Prompt Template 4: Prompt formats used for encounter generation

higher Spearman correlation with the human annotations is selected as the best prompt format and the answers of the LLMs for the particular prompt format were taken as the final judgement of the particular LLM.

Table 1: Statistics obtained by finetuning the different models and the accuracy in obtaining a coherent monster list for a given monster

Fine-tuned Model	No of Epochs	Finetuning Time (Minutes)	Accuracy (%)
Model1v0.1	1	38.0	20.0
Model1v0.2	2	87.0	27.5
Model2v0.1	1	41.0	22.5
Model2v0.2	2	91.0	37.5
Model3v0.1	1	57.0	22.5
Model3v0.2	2	118.0	30.0
Model4v0.1	1	64.0	27.5
Model4v0.2	2	132.0	40.0

4 Results

The results of the model training, prompt engineering, and the analysis done on the outputs from the fine-tuned LLM for the best prompt are provided in this section. Table 1 shows the time that was taken for fine-tuning each of the 8 models and the percentage of correct answers obtained for the 20 frequently used monsters with the basic prompt formats, prompt 1 and prompt 2. It was seen that the models fine-tuned for 1 epoch provided comparatively very inaccurate answers compared to the models trained for 2 epochs. In addition, from the accuracies in the table, it can be seen that adding a role in the instruction dataset has resulted in the increase of the percentage of correct answers. Also, it could be observed that using all the question formats instead of selecting a random one provided slightly better results with the basic prompts. `Model2v0.2` and `Model4v0.2` were taken as the best models and tested with 10 prompt types with 20 frequently used monsters for the next experiment.

Results were obtained for the 2 best models (`Model2v0.2` and `Model4v0.2`) with 10 different types of prompt formats for a set of 20 monsters. From the results obtained, it can be seen that the best result obtained for the `Model4v0.2` was 50.0% which is less compared to the first 2 best results obtained for the `Model2v0.2` which are 70.0% and 55% respectively. Considering the overall results, it can be seen that prompt 9 and prompt 10 with the `Model2v0.2` yielded the best results. And it is noted from Prompt Template 4, that prompt 9 and prompt 10 are based on providing two examples of coherent monster sets in addition to the input

Table 2: Inter-Annotatement agreements between pairs of judges as measured by Spearman correlation. We also show the Spearman Correlation scaled by the Human 1 to Human 2 value to show the relative success of the LLMs.

	Human 1		Human 2		GPT-4.1		Gemini-2.5 flash	
	Raw	Scaled	Raw	Scaled	Raw	Scaled	Raw	Scaled
Human 2	0.49	1.00						
GPT-4.1	0.39	0.80	0.44	0.90				
Gemini-2.5-flash	0.45	0.92	0.53	1.08	0.59	1.20		
DeepSeek-V3-0324	0.48	0.98	0.43	0.88	0.53	1.08	0.53	1.08

Table 3: Percentage of correct, partial and wrong answers for the set of 241 different monsters for the best prompting method according to different judges

Result	Human 1	Human 2	GPT-4.1	Gemini-2.5 flash	DeepSeek-V3 0324	Overall
Coherent monsters in output	37.3	48.1	45.2	37.8	28.6	39.4
Partially coherent monsters in output	21.2	21.2	11.6	31.5	47.7	26.6
Not coherent	41.5	30.7	43.2	30.7	23.7	33.9

query monster and also uses a *role* ("*You are a D&D expert planning to create coherent D&D encounters.*") in the prompt. Conversely, it can be seen when considering the other prompting methods, some commonly used techniques such as asking the LLM to walk through the steps (Prompt 5), asking for explanation (Prompt 4), asking to include the prompt in the answer (Prompt 6) did not work well for this task. From the above results, `Model2v0.2` was considered as the best model and Prompt 9 was taken as the best prompt for the next experiment. These are the final best model and the best prompt used for extensive experimentation. The full results of this experiment is given in Appendix A.

The best prompt was applied to the set of 241 different monsters in D&D and the outputs were obtained. Table 2 shows the Spearman correlations between the different judges for the best judgment prompt. Similar to Palpanadan et al. (2022) and Van Aswegen and Engelbrecht (2009), we use the ranges defined by Guilford (1950) to determine the strength of the correlation. The inter-annotator agreement between the human judges, as measured by Spearman correlation, is observed to be 0.49, which can be considered as a moderate correlation. Therefore, any AI-Human correlation that approaches this value may be taken as reasonable. In order to highlight this relative measure, we have added *scaled* columns to Table 2 where each of the results are scaled as a ratio over the Human-Human correlation value. Considering the best judgment prompt, average AI-AI agreement as measured by Spearman correlation is observed to be 0.54, and the average Human-AI agreement as measured by Spearman correlation is observed to be 0.47, which are also observed to be moderate correlations. When scaled by the Human-Human value, the average Human-AI agreement can be taken as 0.96. The AI-AI agreement, in fact, exceeds 100% when scaled by the Human-Human value.

Further, we conducted a judgment analysis where the results of the best prompt was judged by humans and other LLMs, following the *LLM as Judge* experiment regimen proposed by works such as Gunathilaka and de Silva (2025). Table 3 shows the results obtained by the judgment results for the finetuned LLM by 5 different judges and the overall results. We show the full results of this experiment in Appendix B.

When tested with the best prompt for a set of 241 monsters, some interesting observations were had. For some of the prompts, answers obtained were not wrong but general. Which means they did not specify any monster, but did provide the general name that represented a category of monsters. Also, in some other prompts pertaining to the monsters that are usually found alone, the LLM correctly provided the answer that the monster hunts alone. This by itself is proof that the LLM has correctly learnt the lore. An analysis on whether the output of the LLM is consistent across CR levels is shown in Fig 3. It shows the comparison of cumulative counts of success, partial success, and failure as a percentage across different challenge ratings. The analysis shows that for lower challenge ratings (<1), the cumulative success percentage shows a minor dip. But after CR 1, the trends stabilize. *This*

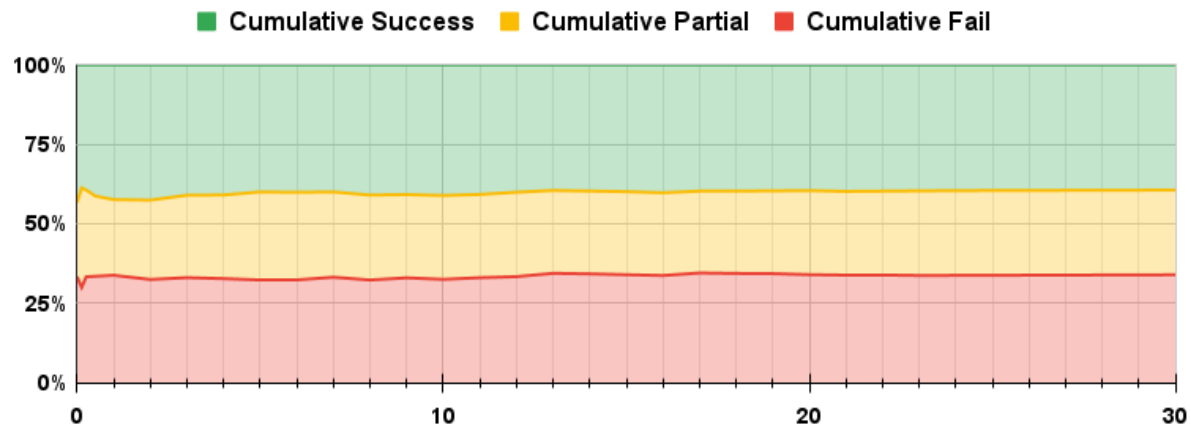

Figure 3: Comparison of cumulative counts of correct, partially correct and failure outputs across Challenge ratings of monsters

may be due to the *boss* and minion format in the prompt. Naturally, when given a monster as a *boss*, the LLM is trying to match them up with potential *minions* who by definition should be of lower power level (CR). But, these monsters being the lowest rung of monsters, the LLM was having a hard time proposing even lower-level monsters to be minions of them.

5 Conclusion and Future Directions

An LLM for generating encounters for a Dungeons and Dragons (D&D) was fine-tuned by instruction tuning an already instruction-tuned `Mistral7B` based LLM using `QLoRa`. The results show that 66.0% of encounters were cohesive or partially cohesive. We can see that further instruction tuning of an already instruction-tuned model is much effective at adapting LLMs to a different domain. The results of the subsequent prompt engineering show that role prompting with examples, which provide the monster combination for an encounter as boss and minions, was the most effective prompt for this particular application. It was also observed that in some instances, in place of the concrete monster name, the LLM provided the category of monsters the result. This may be an artefact of the presence of category level relations in the lore text (eg: "*Mind Flayers* may be seen with other creatures from the *Far Realm*").

Some of the results may also have been affected by the facts that: 1) Not all monsters can be candidates for the boss and minion format, 2) Some monsters are defined as solo creatures in the lore. In cases where these conditions were in play, the LLM would have provided an explanation of the impossibility of creating an encounter. Our current rigid evaluation criteria would have taken such instances as failures on the part of the LLM. However, we consider correcting this to be out of scope for this work and point out that this error results in an under-counting and not an over-counting. Thus, our reported accuracies are a strict lower limit to the actual possible human perceived accuracy. A human DM will find some of the results that we have currently rejected as wrong, to be reasonably acceptable.

As a further future direction, it is planned to augment our system to provide multiple potential encounters in a singular prompt and then, provide a ranking of the encounters based on the coherence to the lore. It is expected that providing the DMs with such a choice may lead to better usability of the system.

References

Justice Ramin Arman, Dan Dillon, and F. Wesley Schneider. 2023. *Planescape: Adventures in the Multiverse*. Wizards of the Coast.

Zahra Ashktorab, Michael Desmond, Qian Pan, James M Johnson, Martin Santillan Cooper, Elizabeth M Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. 2024. Aligning human and llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *arXiv preprint arXiv:2410.00873*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael

Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9379–9393. Association for Computational Linguistics.

A Chowdhery and 1 others. 2022. Scaling language modeling with pathways.

Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. 2024. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*.

Jeremy Crawford, Christopher Perkins, and James Wyatt. 2014a. *Dungeon Master's Guide*. Wizards of the Coast LLC.

Jeremy Crawford, James Wyatt, and Keith Baker. 2019. *Eberron: Rising from the Last War*. Wizards of the Coast.

Jeremy Crawford, James Wyatt, Robert J Schwalb, and Bruce R Cordell. 2014b. *Player's Handbook*. Wizards of the Coast LLC.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 36.

Simon Ellis and James Hendler. 2017. Computers play chess, computers play go… humans play dungeons & dragons. *IEEE Intelligent Systems*, 32(4):31–34.

Haruka Fujiwara, Renta Kimura, and Tokuniki Nakano. 2024. Modify Mistral Large Performance with Low-Rank Adaptation (LoRA) on the BIG-Bench Dataset. *ResearchSquare*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Joy Paul Guilford. 1950. *Fundamental statistics in psychology and education*. McGraw-Hill.

Sadeep Gunathilaka and Nisansa de Silva. 2025. Automatic Analysis of App Reviews Using LLMs. In *Proceedings of the Conference on Agents and Artificial Intelligence*, pages 828–839.

Gary Gygax and Dave Arneson. 1974. *Dungeons & dragons*, volume 19. Tactical Studies Rules Lake Geneva, WI.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Adam Lee, James Introcaso, Ari Levitch, Mike Mearls, Lysa Penrose, Christopher Perkins, Ben Petrisor, Matthew Sernett, Kate Welch, Richard Whitters, and Shawn Wood. 2019. *Baldur's Gate: Descent into Avernus*. Wizards of the Coast.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Annie Louis and Charles Sutton. 2018. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713. Association for Computational Linguistics.

Lara J Martin, Srijan Sood, and Mark O Riedl. 2018. Dungeons and dqns: Toward reinforcement learning agents that play tabletop roleplaying games. In *INTWICED*.

Sarala Thulasi Palpanadan, Toong Hai Sam, Khairunesa Isa, Nurliyana Md Rosni, Asokan Vasudevan, Kai Wah Cheng, and Xue Ruiteng. 2022. Relationship between knowledge level and online consumer purchasing attitude during covid-19 endemic phase. *resmilitaris*, 12(5):352–363.

Akila Peiris and Nisansa de Silva. 2022. Synthesis and evaluation of a domain-specific large data set for dungeons & dragons. *arXiv preprint arXiv:2212.09080*.

Akila Peiris and Nisansa de Silva. 2023. SHADE: Semantic Hypernym Annotator for Domain-Specific Entities-Dungeons and Dragons Domain Use Case. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 1–6. IEEE.

Christopher Perkins, Adam Lee, Richard Whitters, and Jeremy Crawford. 2015. *Curse of Strahd*. Wizards of the Coast.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741.

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

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, and 1 others. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Aravinth Sivaganeshan and Nisansa De Silva. 2023. Fine tuning named entity extraction models for the fantasy domain. In *2023 Moratuwa Engineering Research Conference (MERCon)*, pages 346–351. IEEE.

Kurt Squire. 2007. *Open-ended video games: A model for developing learning for the interactive age*. MacArthur Foundation Digital Media and Learning Initiative.

Eddo Stern. 2002. A touch of medieval: Narrative, magic and computer technology in massively multiplayer computer role-playing games. In *CGDC Conf.*

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A Strong, Replicable Instruction-Following Model.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2024. Are expert-level language models expert-level annotators? *arXiv preprint arXiv:2410.03254*.

Anja S Van Aswegen and Amos S Engelbrecht. 2009. The relationship between transformational leadership, integrity and an ethical climate in organizations. *SA Journal of Human Resource Management*, 7(1):1–9.

Gayashan Weerasundara and Nisansa de Silva. 2023. Comparative analysis of named entity recognition in the dungeons and dragons domain. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1225–1233. INCOMA Ltd., Shoumen, Bulgaria.

Gayashan Weerasundara and Nisansa de Silva. 2024. A Multi-Stage Approach to Image Consistency in Zero-Shot Character Art Generation for the D&D Domain. In *ICAART (3)*, pages 235–242.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Dylan Zhang, Justin Wang, and Francois Charton. 2024. Instruction diversity drives generalization to unseen tasks. *arXiv preprint arXiv:2402.10891*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

A Detailed Results of the prompting experiments

The detailed results of the prompting experiments that were used to compare the best 2 models from the initial experiments and also to select the best prompt format with the best model for the final experiment are given in Table 4. The selection of 20 monsters for this experiment was done manually to cover monsters belonging to different environments, types and alignments to ensure that the testing is unbiased. It can be observed that the combination of `Model2v0.2` and `Prompt 09` yields the best result while specifically for `Model4v0.2`, `Prompt 02` suits best. Overall, most prompts seem to struggle with monsters such as, *Storm Giant*, *Iron Golem*, and *Green Hag* while monsters such as *Lich*, *Sahuagin Baron*, and *Bandit Captain* seem to be easy for most prompts to handle.

B LLM-as-a-Judge Experiments

`GPT-4.1`, `Gemini-2.5-flash` and `DeepSeek-V3-0324` were used with the different judgement prompts to obtain judgements from the relevant LLMs. Prompt Template 5 shows the different prompts that were tried. The judgement prompts were created to impose different types of conditions to the judge the outputs of the fine-tuned LLM. For example, `Judgement Prompt 1` imposes several conditions that a human judges might look for in the fine-tuned LLM's output and `Judgement Prompt 6` simply asks the Judge LLM to rate the fine-tuned LLM's output without imposing any condition.

Results were obtained for the above experiment from each judge (2 human, 3 LLM). Based on this, agreement percentages and spearman correlation were calculated between judgements of each pair of judges to analyze the relationship between the

Table 4: Percentage of correct answers on the initial reference set of 20 monsters for different prompting methods along with basic monster statistics. The Alignment is given as AB where A={L: Lawful, N: Neutral, C: Chaotic} and B={G: Good, N: Neutral, E: Evil}

Prompt	Monster Stats										Prompt Results for Model2v0.2										Prompt Results for Model4v0.2									
	CR	Type	Strength	Dexterity	Constitution	Intelligence	Wisdom	Charisma	Alignment	Habitat	Prompt 01	Prompt 02	Prompt 03	Prompt 04	Prompt 05	Prompt 06	Prompt 07	Prompt 08	Prompt 09	Prompt 10	Prompt 01	Prompt 02	Prompt 03	Prompt 04	Prompt 05	Prompt 06	Prompt 07	Prompt 08	Prompt 09	Prompt 10
Skeleton	1/4	Undead	10	14	15	6	8	5	LE	Urban	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✗
Bandit Captain	2	Humanoid	15	16	14	14	11	14	Any	Artic Coastal Desert Forest Hill Urban	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intellect Devourer	2	Aberration	6	14	13	12	11	20	LE	Underdark	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Hobgoblin Captain	3	Fey	15	14	14	12	10	13	LE	Any	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓
Green Hag	3	Fey	18	12	16	13	14	14	NE	Forest Hill Swamp	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Owlbear	3	Monstrosity	20	12	17	3	12	7	U	Forest	✗	✗	✗	✓	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Water Elemental	5	Elemental	18	14	18	5	10	8	N	Coastal Swamp Underwater	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Sahuagin Baron	5	Fiend	19	15	16	14	13	17	LE	Coastal Underwater	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓
Hill giant	5	Giant	21	8	19	5	9	6	CE	Hill	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✓	✗	✗
Treant	9	Plant	23	8	21	12	16	12	CG	Forest	✗	✗	✓	✗	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓	✗	✗
Aboleth	10	Aberration	21	9	15	18	15	18	LE	Underdark Underwater	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Elder Oblex	10	Ooze	15	16	21	22	17	18	LE	Swamp Underdark Urban	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
Djinni	11	Elemental	21	15	22	15	16	20	CG	Coastal	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗
Beholder	13	Aberration	16	14	18	17	15	17	LE	Underdark	✗	✓	✗	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Storm Giant	13	Giant	29	14	20	16	20	18	CG	Coastal Underwater	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Iron Golem	16	Construct	24	9	20	3	11	1	U	Any	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
Adult blue dragon	16	Dragon	25	10	23	16	15	19	LE	Coastal Desert	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Pit Fiend	20	Fiend	26	14	24	22	18	24	LE	Any	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓
Lich	21	Undead	11	16	16	21	14	16	NE	Any	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓
Solar	21	Celestial	26	22	26	25	25	30	LG	Any	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗
Percentage											20.0	35.0	40.0	35.0	25.0	30.0	40.0	40.0	**70.0**	55.0	20.0	**60.0**	50.0	20.0	25.0	40.0	40.0	40.0	35.0	30.0

Table 5: Percentage of agreement between the judgements of different pairs of judges for different Judgement Prompts

Agreement (%)	Judgment Prompts					
	J-Prompt 1	J-Prompt 2	J-Prompt 3	J-Prompt 4	J-Prompt 5	J-Prompt 6
H1 vs H2	58.92	58.92	58.92	58.92	58.92	58.92
H1 vs GPT	53.94	52.70	50.62	56.43	53.11	55.19
H1 vs Gemini	48.96	51.45	46.47	55.60	51.87	57.68
H1 vs DeepSeek	44.81	49.38	46.06	48.13	46.06	38.59
H2 vs GPT	51.45	49.79	51.45	56.02	57.26	53.11
H2 vs Gemini	55.60	46.47	49.38	54.77	49.38	56.02
H2 vs DeepSeek	49.79	50.62	46.89	47.72	45.64	36.51
GPT vs Gemini	58.92	63.90	58.51	63.07	52.70	61.41
GPT vs DeepSeek	43.15	46.89	39.42	52.70	48.55	49.79
Gemini vs DeepSeek	44.40	48.13	41.49	54.36	54.36	48.13

Table 6: Spearman correlation between the judgements of different pairs of judges for different Judgement Prompts

Spearman Correlation	Judgment Prompts					
	J-Prompt 1	J-Prompt 2	J-Prompt 3	J-Prompt 4	J-Prompt 5	J-Prompt 6
H1 vs H2	0.49	0.49	0.49	0.49	0.49	0.49
H1 vs GPT	0.36	0.37	0.34	0.40	0.37	0.47
H1 vs Gemini	0.36	0.40	0.37	0.45	0.43	0.51
H1 vs DeepSeek	0.29	0.40	0.34	0.48	0.41	0.29
H2 vs GPT	0.33	0.29	0.31	0.44	0.45	0.46
H2 vs Gemini	0.42	0.36	0.39	0.53	0.40	0.49
H2 vs DeepSeek	0.29	0.33	0.27	0.43	0.38	0.18
GPT vs Gemini	0.56	0.57	0.48	0.59	0.51	0.65
GPT vs DeepSeek	0.14	0.24	0.15	0.53	0.42	0.32
Gemini vs DeepSeek	0.25	0.36	0.24	0.53	0.38	0.30

Prompt Formats

1. **Judgement Prompt 1 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. For monsters that hunt alone, it is correct if the output has the monster that is in the question. Answer should be only correct, wrong or partially correct.
2. **Judgement Prompt 2 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. For monsters that hunt alone, it is correct if the output has the monster that is in the question. Answer should be only correct, wrong or partially correct. Be a tough grader.
3. **Judgement Prompt 3 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. For monsters that hunt alone, it is correct if the output has the monster that is in the question. Answer should be only correct, wrong or partially correct. Consider that the ratio of correct:partially correct:wrong answers is 90:51:100.
4. **Judgement Prompt 4 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output without considering the creativity. Answer should be only correct, wrong or partially correct.
5. **Judgement Prompt 5 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong considering only the monster names in the output. Answer should be only correct, wrong or partially correct.
6. **Judgement Prompt 6 :** You are a D&D expert. Given Below is the input given to an LLM and the output generated by the LLM. Judge whether output is correct, partially correct or wrong. Answer should be only correct, wrong or partially correct.

Prompt Template 5: Prompt Formats given to LLMs for obtaining LLM judgements

judgment of different judges on the fine-tuned LLM outputs. Table 5 shows the agreement percentages between each pairs of judges. Table 6 shows the spearman correlations between the different pairs of judges.