A Computational Approach to Generating Lore-Coherent D&D Encounters

229306L - Sivaganeshan Aravinth

Supervised by - Dr. Nisansa de Silva

Content

- Dungeons & Dragons.
- Research Problem.
- Research objectives.
- Literature Survey.
- Extracting Monsters from Fantasy Texts.
- Community detection in D&D monsters.
- Word2Vec for creating encounters.
- LLM fine-tuning and benchmark.
- Discussion


Dungeons & Dragons

Dungeons & Dragons

- Open ended, table-top roleplay game.
- Since 1974 , now in 5th edition [1].
- Predefined rules for gameplay [2-4].
- Several settings.
  - **Lore** for each setting.
  - History, status, relationships, description.
- **Dungeon master (DM)** is the person who conducts the game being the lead story teller.

[1] J. Crawford, J. Wyatt, R. J. Schwalb, and B. R. Cordell, Player’s Handbook. Wizards of the Coast LLC, 2014.
[2] Peiris, A., & de Silva, N. (2022, October). Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation* (pp. 415-424).
[3] K. Squire, Open-ended video games: A model for developing learning for the interactive age. MacArthur Foundation Digital Media and Learning Initiative, 2007.
[4] A. Peiris and N. de Silva, “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). IEEE, 2023, pp. 1–6.

Combat Encounters

- Party of **players**.
- Players face **monsters**.
- The way to attain progress [5].
- Need to **align** with lore.
  - Environment (Immersion [6]):
    - Party going through **forest** attacked by a **tiger**.
    - Party going in a boat in **sea** attacked by a **tiger**.
  - Cohesion (Verisimilitude [6]):
    - The party is attacked by a **herd of sheep** led by a **lioness**.
    - The party is attacked by a **pride of lions** led by a **lioness**.

[5] J. Crawford, C. Perkins, and J. Wyatt, Dungeon Master’s Guide. Wizards of the Coast LLC, 2014.
[6] E. Stern, “A touch of medieval: Narrative, magic and computer technology in massively multiplayer computer role-playing games.” in CGDC Conf., 2002.

Research Problem

Research Problem

"Given a text prompt, creating cohesive D&D encounters with monsters and situation that align with the lore."

- No encounter generator to select monsters automatically according to lore.
- Dungeon masters need to try various combinations.
- Automatic generation remains a gap.

Research Objectives

Research Objectives

R1: To **analyze** existing Dungeons & Dragons datasets and literature on NLP tasks to identify the **information extraction techniques** suitable for automated encounter generation.

R2: To **apply** named entity recognition to **extract monster entities** from D&D texts.

R3: To **construct** association maps **capturing relationships** between these entities, enabling automated, lore-consistent encounter generation.

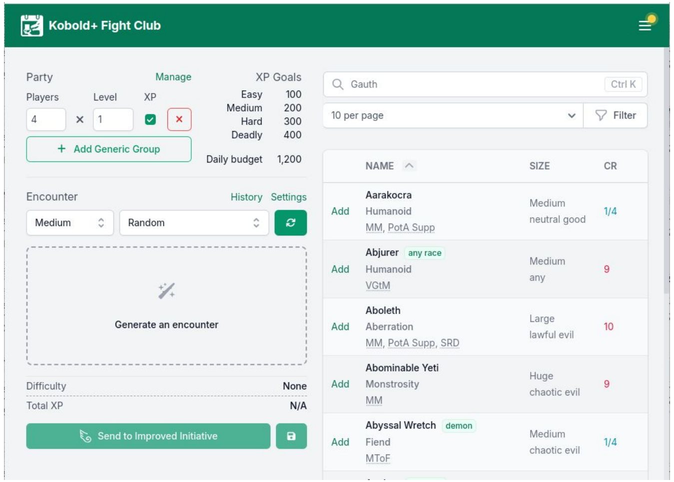
R4: To **design** a system for **automated selection of monsters** for D&D encounters

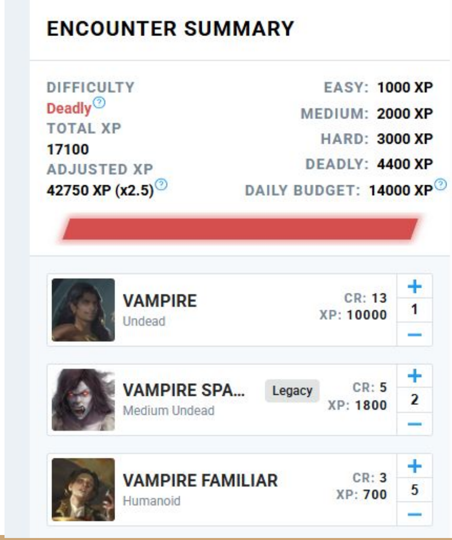
R5: To **validate** the generated encounters **for coherence with the lore**.

Literature Survey

Encounter Generation

- Publicly available online tools[1][2][3][4].
    - Using challenge rating system
    - Calculates XP point thresholds for player level inputs.
    - When a player selects monsters, calculates XP points for the monster combination.
    - Based on threshold, outputs the difficulty.





1 https://www.dndbeyond.com/encounter-builder
2 https://www.aidedd.org/dnd-encounter/
3 https://www.kassoon.com/dnd/5e/generate-encounter/
4 https://koboldplus.club/

Existing Works on D&D

- Forgotten Realms Wiki (`FRW`) dataset [2].
  - 11 datasets from FRW.
  - Free text generator.
  - Named entity classifier.
- Studies using D&D Data.
  - Dialogue and discourse analysis [7].
  - AI playing the game as DMs or players [8].
  - NER [9].
  - Image and Adventure generation [2,10].

[2] Peiris, A., & de Silva, N. (2022, October). Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation* (pp. 415-424).
[7] 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.
[8] Simon Ellis and James Hendler. 2017. Computers play chess, computers play go. . . humans play dungeons & dragons. IEEE Intelligent Systems, 32(4):31–34.
[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.
[10] Gayashan Weerasundara and Nisansa de Silva. 2024. A Multi-Stage Approach to Image Consistency in ZeroShot Character Art Generation for the D&D Domain. In ICAART (3), pages 235–242.

Named Entity Recognition

- Models
  - LSTM, CNN and CRF based models.
    - Bi-LSTM-CRF [11] , Bi-LSTM-Bi-CRF [12].
    - Bi-LSTM-CNN [13].
  - Contextualized string representations [14].
  - Pre trained transformer based models.
    - BERT [15] with sequence tagging layer.
    - FLERT [16], Luoma and Pyysalo [17].

- Frameworks
  - Trankit [18].
  - Stanza [19].
    - Contextualised string representation model.
  - FLAIR [20].
    - Contextualised string representation model.
    - FLERT.
  - spaCy[4].

[11] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.
[12] R. Panchendrarajan and A. Amaresan, "Bidirectional lstm-crf for named entity recognition," in Proceedings of the 32nd Pacific Asia conference on language, information and computation, 2018.
[13] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," Transactions of the association for computational linguistics, vol. 4, pp.357–370, 2016.
[14] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1638–1649.
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[16] S. Schweter and A. Akbik, "Flert: Document-level features for named entity recognition," arXiv preprint arXiv:2011.06993, 2020.
[17] J. Luoma and S. Pyysalo, "Exploring cross-sentence contexts for named entity recognition with bert," in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 904–914.
[18] M. Van Nguyen, V. D. Lai, A. P. B. Veyseh, and T. H. Nguyen, "Trankit: A light-weight transformerbased toolkit for multilingual natural language processing," arXiv preprint arXiv:2101.03289, 2021.
[19] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," arXiv preprint arXiv:2003.07082, 2020.
[20] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.

4 https://spacy.io/

Coreference Resolution

- End to end deep learning models.
  - Optimizing the margin likelihood of antecedent spans [21].
  - Iterative span ranking architecture [22].
- Language representation models.
  - BERT for span ranking [23].

[21] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, “End-to-end neural coreference resolution,” arXiv preprint arXiv:1707.07045, 2017.
[22] K. Lee, L. He, and L. Zettlemoyer, “Higher-order coreference resolution with coarse-to-fine inference,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 687–692. [Online]. Available: https://aclanthology.org/N18-2108
[23] M. Joshi, O. Levy, D. S. Weld, and L. Zettlemoyer, “Bert for coreference resolution: Baselines and analysis,” arXiv preprint arXiv:1908.09091, 2019.

Word Embeddings

- Vector representations for words.
  - One hot encoding.
  - Bag of words.
- Word2Vec [24].
  - CBOW.
  - Skip gram embeddings.
- GloVe [25].
- FastText [26].

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
[25] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
[26] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

Community Detection

- Traditional methods.
  - Modularity based methods [27,28].
  - Label propagation based methods [29].
- Deep learning based methods.
  - Communitygan [30].
  - Non-negative Matrix Factorization [31,32].
  - Graph neural networks [33,34].

[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.
[28] V. A. Traag, “Faster unfolding of communities: Speeding up the louvain algorithm,” Physical Review E, vol. 92, no. 3, p. 032801, 2015.
[29] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” Physical review E, vol. 76, no. 3, p. 036106, 2007.
[30] Y. Jia, Q. Zhang, W. Zhang, and X. Wang, “Communitygan: Community detection with generative adversarial nets,” in The World Wide Web Conference, 2019, pp. 784–794.
[31] Y. Li, C. Sha, X. Huang, and Y. Zhang, “Community detection in attributed graphs: An embedding approach,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.
[32] F. Ye, C. Chen, and Z. Zheng, “Deep autoencoder-like nonnegative matrix factorization for community detection,” in Proceedings of the 27th ACM international conference on information and knowledge management, 2018, pp. 1393–1402.
[33] Z. Chen, X. Li, and J. Bruna, “Supervised community detection with line graph neural networks,” arXiv preprint arXiv:1705.08415, 2017.
[34] D. Jin, Z. Liu, W. Li, D. He, and W. Zhang, “Graph convolutional networks meet markov random fields: Semi-supervised community detection in attribute networks,” in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 152–159.

Fine-tuning and Using LLMs

- Fine-tuning LLMs
  - Instruction tuning [35].
    - Adapting without extensive retraining or architectural changes [35].
    - Alpaca dataset format is a standard structure for instruction tuning [37].
  - Low Rank Adaptation [38].
    - Useful for fine-tuning large models to downstream tasks without updating all the parameters of the existing model.
    - QLora [36] - Fine-tuning by quantizing the pretrained model to 4 bits.
- Using LLMs
  - Prompt Engineering.
    - Better prompts lead to better outputs [39].
    - Main categories of In-Context Learning, thought generation, decomposition, ensembling, and self-criticism [40].
  - LLM-as-a-Judge.
    - Provides scalable, flexible and consistent assessments [41].
    - Replace human judges or used together with human judges for rapid, scalable evaluation [42].

[35] 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.
[36] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient fine tuning of quantized llms. NeurIPS, 36.
[37] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... & Hashimoto, T. B. (2023). Alpaca: a strong, replicable instruction-following model; 2023. URL https://crfm. stanford. edu/2023/03/13/alpaca. html.
[38] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.
[39] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
[40] Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... & Resnik, P. (2024). The prompt report: a systematic survey of prompt engineering techniques. arXiv preprint arXiv:2406.06608.
[41] 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.
[42]Ashktorab, Z., Desmond, M., Pan, Q., Johnson, J. M., Cooper, M. S., Daly, E. M., ... & Geyer, W. (2024). Aligning human and llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. arXiv preprint arXiv:2410.00873.

Extracting Monsters from Fantasy Texts

Data

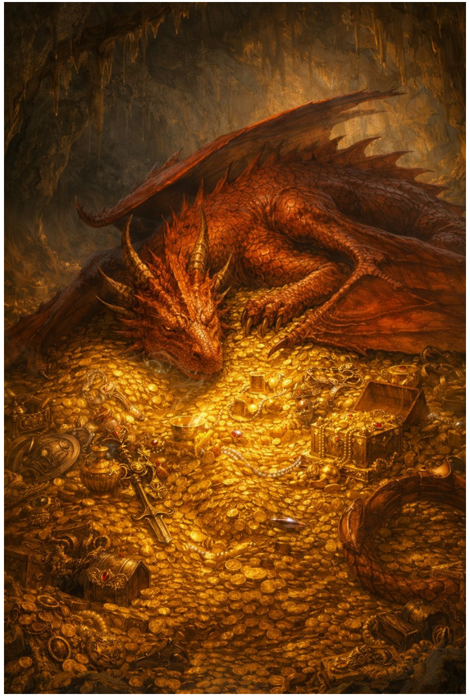
- Need to have several mentions of monster entities.
- Monster list
  - Publicly available monster list (D&D) **System Reference Document (SRD)** monster list[2].
  - FRW-I dataset [2].
- Text
  - Publicly available monster lore[3].

A Gold Dragon's Lair

Gold dragons make their homes in out-of-the-way places, where they can do as they please without arousing suspicion or fear. Most dwell near idyllic lakes and rivers, mist-shrouded islands, cave complexes hidden behind sparkling waterfalls, or ancient ruins.

Lair Actions

On initiative count 20 (losing initiative ties), the dragon takes a lair action to cause one of the following effects; the dragon can't use the same effect two rounds in a row.



2 https://5thsrd.org/gamemaster_rules/monster_indexes/monsters_by_cr/

3 https://www.dndbeyond.com/

[2] Peiris, A., & de Silva, N. (2022, October). Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation (pp. 415-424).

Methodology

- Obtaining monster lists.
  - SRD monster list (**317** monsters).
  - FRW-I extracted monster list (**753** monsters).
  - Two lists for experiments: SRD list, merged list (**853** monsters).
- Obtaining text.
  - Publicly available lore text for **246** monsters.
  - Lore data of all **246** monsters combined.
- Setups for creating training, development and test sets.
  - **Setup 1 -** Tagging combined lore data using text lookup for SRD monster list.
  - **Setup 2 -** Tagging combined lore data using text lookup for merged list.

Initial Text Lookup

- Initial text lookup of monsters in the lore data.
- False detections possible (**impress** → **imp**)
- Avoiding false detection.
  - An association map is formed by listing monsters found in each monster’s lore.
  - A threshold of 30 is taken and monsters that appear in more than 30 other monster’s lores are added to a list called **ignore list**.
  - This is done for both setup 1 and setup 2 and respective monsters lists were refined by removing the ignore list.

Text lookup and BIO Tagging

- Combined monster lore was tokenized to sentences by Trankit tokenizer.
- Refined monster list was looked up and the position index of the words in the sentences were stored and sentences were tokenized.
- Using previous position index each of the words were tagged with BIO tags.
    - `B-MONS.`
    - `I-MONS.`
    - `O`
- Same method is followed for both the setup 1 and setup 2.

Data Preparation

- BIO tagged data separated into training, development and test sets.
  - Done without separating sentences from same lore into 2 sets.
  - ⅔ (66.67 %) training set, ⅙ (16.67 %) development set and ⅙ (16.67 %) test.
  - Total 4520 sentences divided into 3011, 764 and 745 sentences.
- The test data that is produced during the process in Setup 2 is manually refined to produce gold standard test data.

```
gold B-MONS
dragons I-MONS
make O
their O
homes O
in O
out O
- O
of O
- O
the O
- O
way O
places O
```

Training and Evaluation

- Zero-shot Trankit (`XLM-RoBerta`[14]) on the monster lore data.
- Models trained and tested with tagged data setup 1 and setup 2.
  - XLM-Roberta pretrained model in `Trankit`.
  - Stacked embeddings as proposed by Akbik et al. [15] available in `FLAIR` (Config 1)
  - Stack embeddings with BiLSTM-CRF Sequence tagger (Config 2).
- Hyper parameters tuned.
  - Batch size of 32 identified for `FLAIR`.
  - Batch size of 16 for Trankit.
  - Number of epochs 25.
- Combined lore data is tagged with the best model identified and association maps and digraphs were created for visualization.

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, ´ and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” arXiv preprint arXiv:1911.02116, 2019.
[15] A. Akbik, D. Blythe, and R. Vollgraf, “Contextual string embeddings for sequence labeling,” in Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1638–1649.

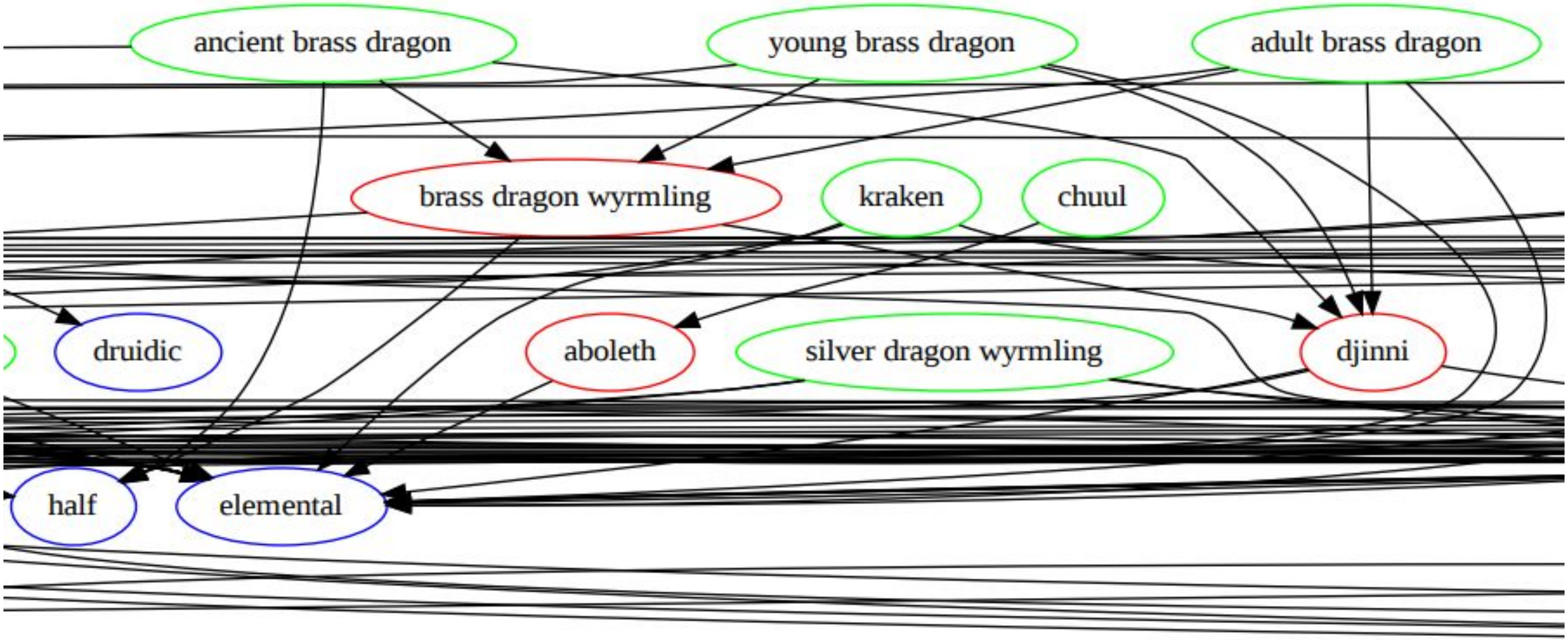
Results - Model Scores

Model	Training	Precision		Recall		F1	
		Text Lookup test data	Manually tagged test data	Text Lookup test data	Manually tagged test data	Text Lookup test data	Manually tagged test data
FLAIR Config 1	Setup 1	62.11	82.97	72.46	28.23	66.89	42.12
	Setup 2	91.91	85.58	92.46	89.17	92.19	87.34
FLAIR Config 2	Setup 1	64.91	86.10	80.43	28.58	71.84	42.87
	Setup 2	91.11	85.47	**92.87**	**89.50**	92.97	87.43
Trankit	Zero shot	0.00	0.00	0.00	0.00	0.00	0.00
	Setup 1	66.42	82.58	65.94	26.78	66.18	40.44
	Setup 2	**96.67**	**86.44**	92.26	89.33	**94.42**	**87.86**

Results - Association Map

```
Text Lookup ancient blue dragon:['blue dragon', 'devil', 'air elemental', 'verdant',
                                  'brass dragon', 'bard', 'wizard', 'assassin', 'ankheg', 'giant scorpion', 'noble']
Trankit tag ancient blue dragon:['blue dragon', 'dust devil', 'air elemental', 'verdant',
                                  'brass dragon', 'minion', 'bard', 'wizard', 'assassin', 'ankheg', 'giant scorpion', 'noble']
Text Lookup ice devil:['ice devil', 'devil', 'pit fiend']
Trankit tag ice devil:['ice devil', 'devil', 'fiend', 'archdevil']
Text Lookup blood hawk:['blood hawk']
Trankit tag blood hawk:['blood hawk']
Text Lookup cultist:['cultist', 'elemental', 'acolyte']
Trankit tag cultist:['cultist', 'elemental', 'acolyte']
Text Lookup bandit:['bandit', 'thug', 'veteran', 'banditry']
Trankit tag bandit:['bandit', 'thug', 'veteran', 'banditry', 'pirate']
Text Lookup flying snake:['flying snake', 'cultist']
Trankit tag flying snake:['flying snake', 'cultist']
```

Results - Digraphs



Discussion

- Zero shot `Trankit` was not able to identify any monster entities.
- `FLAIR` model:
  - LSTM-CRF sequence tagger outperformed the one LSTM tagger.
- Fine tuned `Trankit` outperformed the fine tuned `FLAIR` models:
  - Even some monsters which were not in text lookup were identified by model.
    eg: **Archdevil**, **Dust devil**
  - Some entity names that represent a particular monster according to the context they appear are identified as monsters.
    eg: **The fanatic** in context means **Cult Fanatic**
  - Plurals of fantasy domain specific monsters that cannot be identified by text lookup identified correctly.
    eg: **Djinni** vs **Djinn**

Publications

- *Fine Tuning Named Entity Extraction Models for the Fantasy Domain*
  - Published in MERCON 2023.
  - Includes the comparison of fine-tuning of results different NER models to D&D data.
- Research Objectives Achieved:
  - **R1:** To **analyze** existing Dungeons & Dragons datasets and literature on NLP tasks to identify the **information extraction techniques** suitable for automated encounter generation.
  - **R2:** To **apply** named entity recognition to **extract monster entities** from D&D texts.
  - **R3:** To **construct** association maps **capturing relationships** between these entities, enabling automated, lore-consistent encounter generation.


Embedding & Community Detection

Community Detection

- Louvain [27] method to form environment based communities.
  - Merging of communities.
  - Enriching communities.
- **Result:** 60 multi monster communities obtained and 290 single monster communities.
  - **72%** of multi monster communities were coherent.
  - Considering all the multi-monster communities and the encounters that were obtained in the same monster space, it could be said that **87.7 %** of the encounters were same as or a subset of the monster communities.

[27] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.

Word2Vec on FRW Data

- **Pre-training** architectures used.
  - Continuous Bag-of-Words (C-BOW).
  - Skipgram.
- Obtained models which could identify some D&D relationships:
  - **Lore: Vhaeraun** is the son of **Lolth** and **Glasya** is the daughter of **Asmodeus**.
  - **Evaluate: Asmodeus - Lolth + Vhaeraun**
  - **Expected answer: Glasya**.
  - **Continued Fine-tuning:** For each rule of the form A, B → C, D, adjust the vector of C according to the equation C = A - B + D.
  - **Result:** Success in 72 out of 96 models (75 %).
- The values show that the rules fitting **does not provide expected results** even for the monsters that were in the fitted rules, considering another relationship.
- The applied retrofitting strategy is **not sufficiently reliable** for downstream tasks such as encounter generation.

LLM Fine-tuning and Benchmark

Datasets

- 41,106 instructions were obtained from FRW-J-Alpaca [2].
- Let's call it as ***FRW-I Instruction Dataset***.
- Sources used to create datasets.
  - 5etools[1].
  - A total of 3786 human expert-created encounters were extracted from publicly available data from numerous online sources.

[2] Peiris, A., & de Silva, N. (2022, October). Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation (pp. 415-424).
1 https://5e.tools/

Datasets Created from 5etools

- Linguistic diversity of instructions is seen to help models generalize better [43].
- We formed questions in 3-5 linguistically different formats for each column.
- Irrelevant features dropped.
- Two Datasets created:
  - ***5et-I Instruction Dataset*** - Selecting one format for each monster, feature pair.
  - ***5et-I-All Instruction Dataset*** - Selecting all formats for each monster, feature pair.

[43] Zhang, D., Wang, J., & Charton, F. (2024). Instruction diversity drives generalization to unseen tasks. arXiv preprint arXiv:2402.10891.

Datasets Created from Encounters

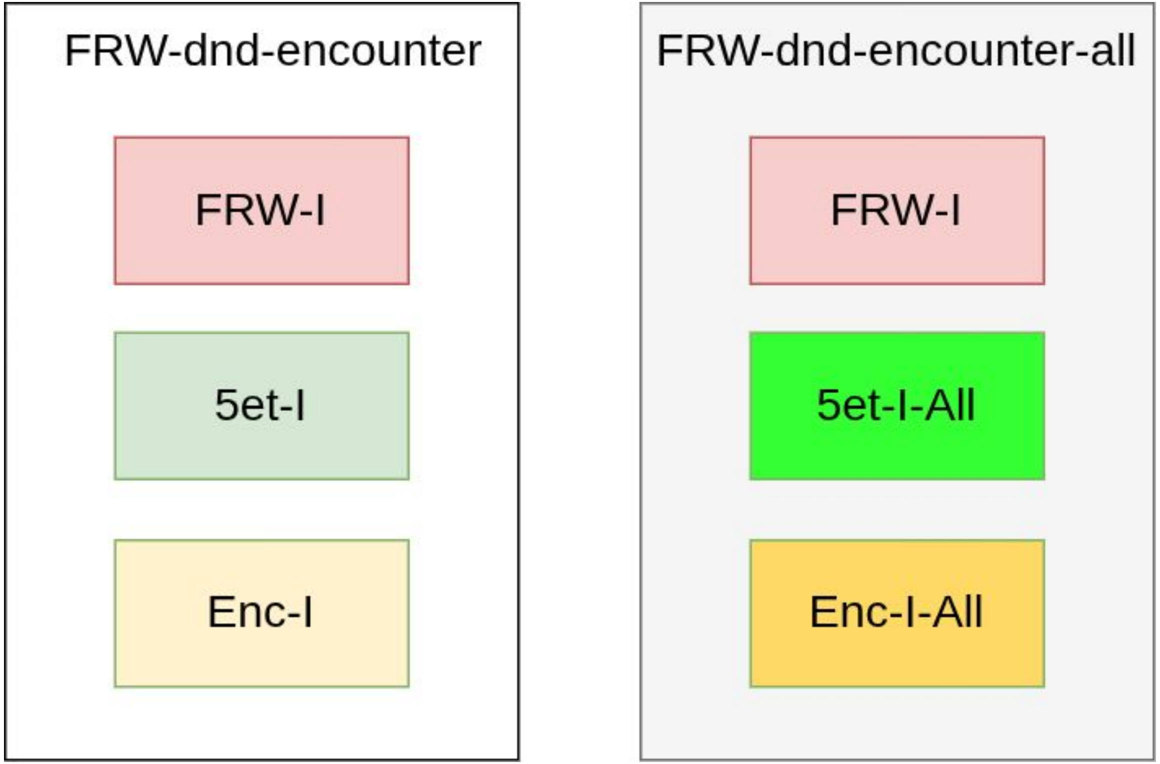
- **Input:** A total of 3786 human expert-created encounters.
- **Output:** 2587 monster *itemsets* that can come together in an encounter were obtained.
- Frequent *itemset* mining using `Apriori` [44].
- Generating linguistically different question formats.
- Two Datasets created:
  - ***Enc-I Instruction Dataset*** - Selecting one question format for each rule.
  - ***Enc-I-All Instruction Dataset -*** Using all formats for each rule.

Enc-I

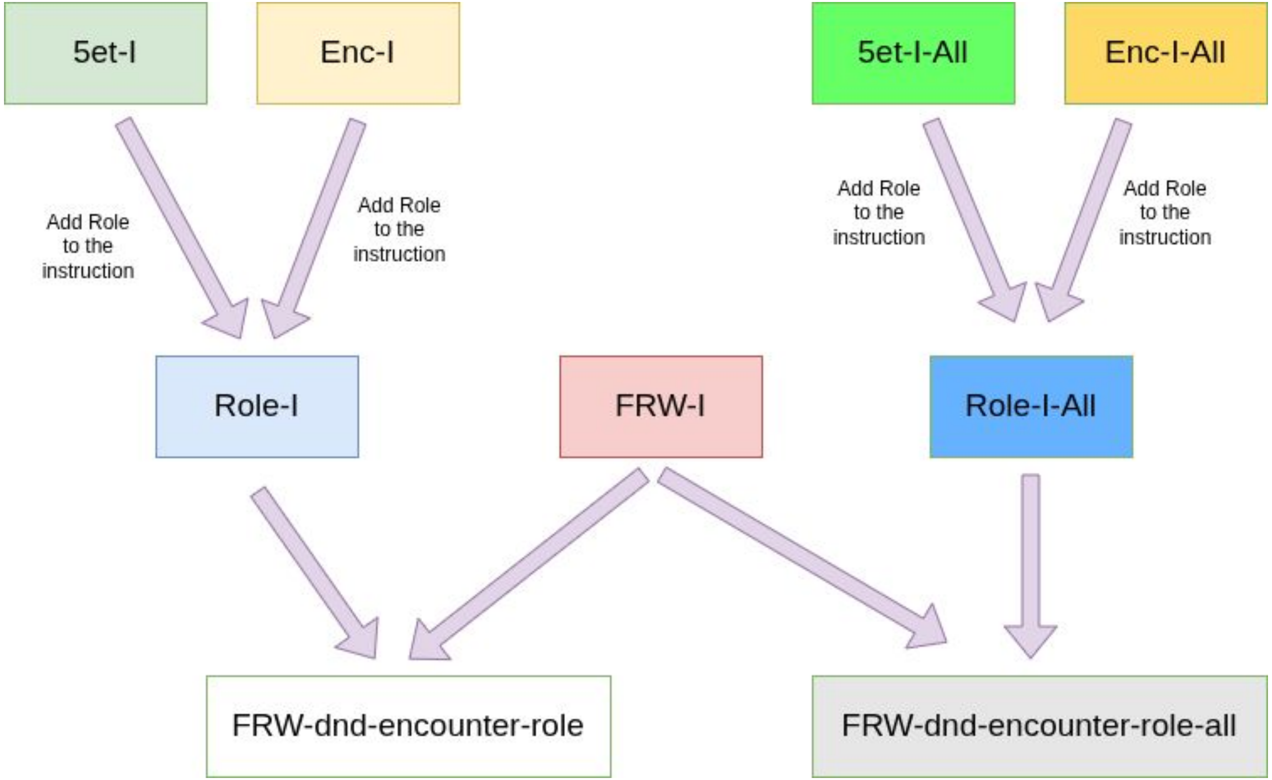
Enc-I-All

[44] Srikant, R. (1996). Fast algorithms for mining association rules and sequential patterns. The University of Wisconsin-Madison.

Final Datasets Used for Instruction Tuning



Final Datasets Used for Instruction Tuning



Fine-tuning the Model

- **Base model :-** *Mistral7BInstructv0.2*[3]
- With 7B parameters, it is shown to be efficient and having comparable performance to the state of the art 13B parameter chat models and also known to work well LORA [45].
- Quantized LORA (QLORA) [36] is utilized to reduce the computing resources used for training.
- 8 models created by instruction tuning with the 4 datasets for 1 and 2 epochs respectively.
- The training was done using *Axolotl* framework in an H100 2XM GPU cloud virtual machine from *runpod.io* with 80 GB VRAM.

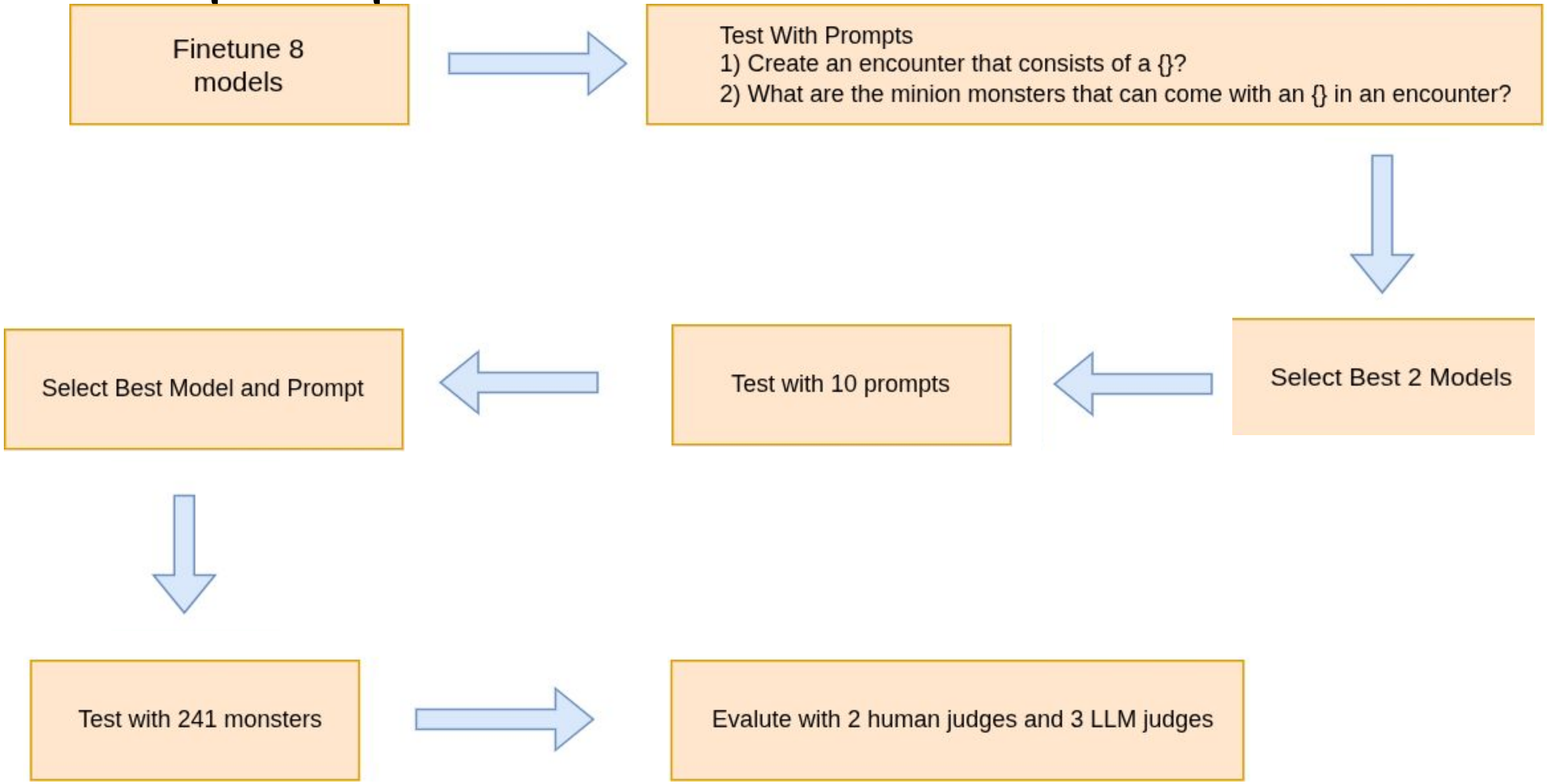
[45] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guil-laume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.
[36] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient fine tuning of quantized llms. NeurIPS, 36.
3 https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Prompt Experiments

Finetune 8 models → Test With Prompts
1) Create an encounter that consists of a {}?
2) What are the minion monsters that can come with an {} in an encounter?

Model	Instruction Dataset	Epochs	Time (Minutes)	Accuracy (%)
Model1v0.1	FRW-dnd-encounter	1	38	20.0
Model1v0.2	FRW-dnd-encounter	2	87	27.5
Model2v0.1	FRW-dnd-encounter-role	1	41	22.5
Model2v0.2	**FRW-dnd-encounter-role**	**2**	**91**	**37.5**
Model3v0.1	FRW-dnd-encounter-all	1	57	22.5
Model3v0.2	FRW-dnd-encounter-all	2	118	30.0
Model4v0.1	FRW-dnd-encounter-role-all	1	64	27.5
Model4v0.2	**FRW-dnd-encounter-role-all**	**2**	**132**	**40.0**

Prompt Experiments



Results for the Best 2 Models

- Model, prompt pair which gave the best results was ***Model2v0.2*** and the ***In Context Learning based prompt***. [40].
  - Two examples of boss monsters and the respective minion monsters provided (Two-Shot)
  - Asking the model to provide the minion monsters for boss monster worked well.
- Some other common prompting techniques [40] did not work well for this task.
  - Asking the LLM to walk through the steps.
  - Asking to include the prompt in the answer.

[40] Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... & Resnik, P. (2024). The prompt report: a systematic survey of prompt engineering techniques. arXiv preprint arXiv:2406.06608.

Results for Best Model with the Best Prompt

- Spearman Correlation

	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

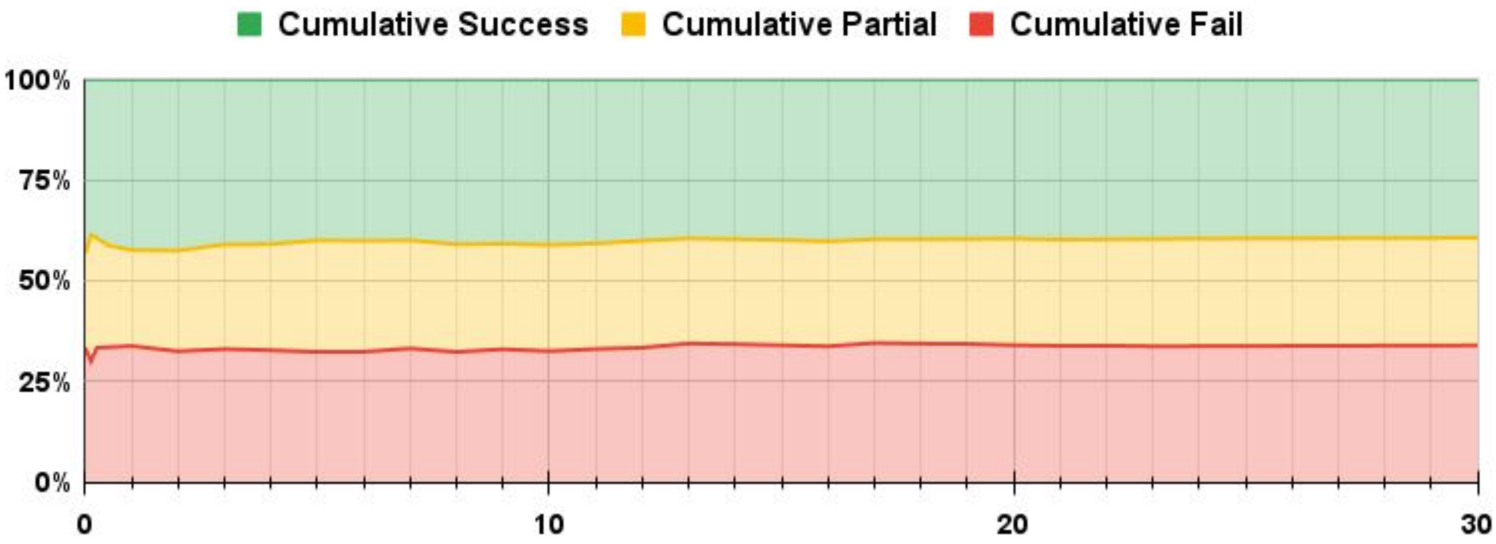
Results for Best Model with the Best Prompt

- 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

Consistency with Monster CR

- Comparison of cumulative counts of correct, partially correct and failure outputs across **Challenge Ratings (CR)** of monsters.



Outputs of the Encounter Generating LLM

- Dip in the correct % in certain lower CR levels as they cannot be a boss monster.
- 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 one point to say that the LLM has correctly learnt the lore.
- Another aspect of wrong answers were that in some cases, the answers were not wrong but, answers had the generic category of the monster which resulted in rejecting that encounter.

Publications

- *Lore Coherent Encounter Generation for Dungeons and Dragons: LLM Fine-tuning and Benchmark*.
  - Published in PACLIC 2025.
  - It includes the methodology followed for fine-tuning an LLM for D&D domain and the results.
- Research Objectives Achieved:
  - **R4:** To **design** a system for **automated selection of monsters** for D&D encounters
  - **R5:** To **validate** the generated encounters **for coherence with the lore**.

Overall Discussion

- Encounter generation.
  - With **66.6%** encounters being fully or partially coherent.
  - Evaluated by 2 Human evaluators and 3 LLMs.
  - Encounters for general category rather than specific monsters in some cases.
  - Encounter according to special characteristics of monsters in some cases.
- Research Contributions.
  - **2 Publications**:
    - In PACLIC 2025 published paper titled “Lore Coherent Encounter Generation for Dungeons and Dragons: LLM Fine-tuning and Benchmark”. (**R1**, **R2**, and **R3**)
    - In MERCON 2023 published paper titled “Fine Tuning Named Entity Extraction Models for the Fantasy Domain”. (**R4** and **R5**)
  - **4 datasets** made available in Huggingface (https://huggingface.co/datasets/Aravinth92/FRW-J_monster_encounter_data/)
  - **8 fine-tuned LLMs** made available in Huggingface (https://huggingface.co/Aravinth92/models)

References

References

1. J. Crawford, J. Wyatt, R. J. Schwalb, and B. R. Cordell, Player's Handbook. Wizards of the Coast LLC, 2014.
2. Peiris, A., & de Silva, N. (2022, October). Synthesis and Evaluation of a Domain-specific Large Data Set for Dungeons & Dragons. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation* (pp. 415-424).
3. K. Squire, Open-ended video games: A model for developing learning for the interactive age. MacArthur Foundation Digital Media and Learning Initiative, 2007.
4. A. Peiris and N. de Silva, "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). IEEE, 2023, pp. 1–6.
5. J. Crawford, C. Perkins, and J. Wyatt, Dungeon Master's Guide. Wizards of the Coast LLC, 2014.
6. E. Stern, "A touch of medieval: Narrative, magic and computer technology in massively multiplayer computer role-playing games." in CGDC Conf., 2002.
7. 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.
8. Simon Ellis and James Hendler. 2017. Computers play chess, computers play go. . . humans play dungeons & dragons. IEEE Intelligent Systems, 32(4):31–34.
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.
10. Gayashan Weerasundara and Nisansa de Silva. 2024. A Multi-Stage Approach to Image Consistency in ZeroShot Character Art Generation for the D&D Domain. In ICAART (3), pages 235–242.
11. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.
12. R. Panchendrarajan and A. Amaresan, "Bidirectional lstm-crf for named entity recognition," in Proceedings of the 32nd Pacific Asia conference on language, information and computation, 2018.
13. J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," Transactions of the association for computational linguistics, vol. 4, pp.357–370, 2016.
14. A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1638–1649.

References

15. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
16. S. Schweter and A. Akbik, “Flert: Document-level features for named entity recognition,” arXiv preprint arXiv:2011.06993, 2020.
17. J. Luoma and S. Pyysalo, “Exploring cross-sentence contexts for named entity recognition with bert,” in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 904–914.
18. M. Van Nguyen, V. D. Lai, A. P. B. Veyseh, and T. H. Nguyen, “Trankit: A light-weight transformerbased toolkit for multilingual natural language processing,” arXiv preprint arXiv:2101.03289, 2021.
19. P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A python natural language processing toolkit for many human languages,” arXiv preprint arXiv:2003.07082, 2020.
20. A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “Flair: An easy-to-use framework for state-of-the-art nlp,” in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), 2019, pp. 54–59.
21. K. Lee, L. He, M. Lewis, and L. Zettlemoyer, “End-to-end neural coreference resolution,” arXiv preprint arXiv:1707.07045, 2017.
22. K. Lee, L. He, and L. Zettlemoyer, “Higher-order coreference resolution with coarse-to-fine inference,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 687–692. [Online]. Available: https://aclanthology.org/N18-2108
23. M. Joshi, O. Levy, D. S. Weld, and L. Zettlemoyer, “Bert for coreference resolution: Baselines and analysis,” arXiv preprint arXiv:1908.09091, 2019.
24. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
25. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
26. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

References

27. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” Journal of statistical mechanics: theory and experiment, vol. 2008, no. 10, p. P10008, 2008.
28. V. A. Traag, “Faster unfolding of communities: Speeding up the louvain algorithm,” Physical Review E, vol. 92, no. 3, p. 032801, 2015.
29. U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” Physical review E, vol. 76, no. 3, p. 036106, 2007.
30. Y. Jia, Q. Zhang, W. Zhang, and X. Wang, “Communitygan: Community detection with generative adversarial nets,” in The World Wide Web Conference, 2019, pp. 784–794.
31. Y. Li, C. Sha, X. Huang, and Y. Zhang, “Community detection in attributed graphs: An embedding approach,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.
32. F. Ye, C. Chen, and Z. Zheng, “Deep autoencoder-like nonnegative matrix factorization for community detection,” in Proceedings of the 27th ACM international conference on information and knowledge management, 2018, pp. 1393–1402.
33. Z. Chen, X. Li, and J. Bruna, “Supervised community detection with line graph neural networks,” arXiv preprint arXiv:1705.08415, 2017.
34. D. Jin, Z. Liu, W. Li, D. He, and W. Zhang, “Graph convolutional networks meet markov random fields: Semi-supervised community detection in attribute networks,” in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 152–159.
35. 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.
36. Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient fine tuning of quantized llms. NeurIPS, 36.
37. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... & Hashimoto, T. B. (2023). Alpaca: a strong, replicable instruction-following model; 2023. URL https://crfm. stanford. edu/2023/03/13/alpaca. html.
38. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.
39. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
40. Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... & Resnik, P. (2024). The prompt report: a systematic survey of prompt engineering techniques. arXiv preprint arXiv:2406.06608.
41. 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.
42. Ashktorab, Z., Desmond, M., Pan, Q., Johnson, J. M., Cooper, M. S., Daly, E. M., ... & Geyer, W. (2024). Aligning human and llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. arXiv preprint arXiv:2410.00873.
43. Zhang, D., Wang, J., & Charton, F. (2024). Instruction diversity drives generalization to unseen tasks. arXiv preprint arXiv:2402.10891.
44. Srikant, R. (1996). Fast algorithms for mining association rules and sequential patterns. The University of Wisconsin-Madison.
45. Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guil-laume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Thank you.

Q/A