Hylian NER: Annotating and Modeling Entities in a Game Domain

Yangyang Chen
Brandeis University
yangyangchen@brandeis.edu

Zixin Jiang
Brandeis University
zixinjiang@brandeis.edu

Abstract

We present a new Named Entity Recognition (NER) system designed for the fictional universe of *The Legend of Zelda: Breath of the Wild*, which recognizes four types of entities: characters, locations, creatures, and items. Since no existing corpus or model targets this game series, we construct a novel dataset from in-game text together with using gazetteers and synthetic augmentation. We annotate a set of 1840 sentences using a large language model (LLM) with carefully constructed prompting techniques. We evaluated our results on an RNN baseline with BiLSTM, a BiLSTM+CRF variant, the BERT model, and the DistilBERT model. The best model achieves an F1 score at the entity level of 0.86, with decent performance across all types of entities. We also attempt fine-tuning several LLMs but are unable to obtain usable output for evaluation. Our results highlight the viability of transformer models for NER of fictional domains. Although LLMs provide promising annotation assistance, they still present a big challenge and appear suboptimal for our NER task.

1 Introduction

Our class project aims at developing a Named Entity Recognition(NER) system tailored to one of the more recent works from The Legend of Zelda series, Breath of the Wild (BOTW). Although most of the NER research focuses on "real-life" domains such as biomedical or newswire data, the fictional domain remains relatively understudied, hense lower-resourced. BOTW was first released by Nintendo exclusively on their Switch console in 2017. This is an extremely narrative-rich game and thus contains some unique entities such as characters (CHAR), locations (LOCA), creatures (CREA), and items (ITEM) in a structured way. Therefore, questions arise such as whether these entities are capable of modeling.

Considering the era with rapid development of large language models (LLMs), another motivation is to explore whether LLMs could help with 1) speeding up annotation; and 2) NER from a characteristic gaming domain.

To our knowledge, no existing corpus or model supports this task. We address this gap by first building a new annotated dataset specific to BOTW and integrating both curated and synthetic data. Then we fine-tuned a couple of language models of different scales – from recurrent neural networks (RNN), to transformer-based models, and finally LLMs. Our project combines domain-specific knowledge extraction with practical NER modeling to investigate how well modern systems can adapt to a video game domain.

Our main contributions include: 1) the creation of a BOTW-specific NER dataset annotated for four entity types (CHAR, LOC, CREA, ITEM) with 1840 sentences; 2) a novel pipeline combining gazetteer-based filtering, synthetic data generation from in-game descriptions, and LLM-assisted annotation. 3) a benchmark comparison of RNN and transformer-based models, with a best F1 score of 0.86; 4) extensive attempt on LLM fine-tuning.

All code and data for this project are available at our GitHub repository.[1]

2 Related Work

NER has been extensively studied in standard domains like news (CoNLL-2003) (Tjong Kim Sang and De Meulder, 2003), biomedical texts (BC5CDR) (Li et al., 2016), and other more "real-life" scenarios. Transformer models like BERT (Devlin et al., 2019) and its variants have set state-of-the-art (SOTA) benchmarks on many of these datasets. However, fictional domains such as video games, movies, or literature remain underexplored.

[1] https://github.com/1192119703jzx/Hylian_NER_for_The_Legend_of_Zelda

Recent work has explored synthetic data augmentation for low-resource NER (Dai and Adel, 2020), and others have used gazetteers to build silver-standard corpora (Mayhew et al., 2019). Sequence tagging with CRF heads over contextual embeddings (e.g., BERT+CRF) has been shown to improve label consistency (Ma and Hovy, 2016).

A few efforts are made specifically for the fictional domains like fictions (Chu et al., 2020), fantasy (Weerasundara and de Silva, 2023), and online games (Liu et al., 2020)

To our knowledge, no prior work builds or evaluates NER models towards the worldview of Zelda or any comparable game universe, indicating that our project is both novel and practically useful for narrative-driven applications.

3 Data

We collect our data mainly from two repositories created by game fans. After significant effort on cleaning and processing, we were able to obtain a relatively well-represented dataset with a size of 1840. One challenge during collection is to handle massive junk data. We have to spend a long time examining the usable data that spread out across 73,294 lines of raw text dump. Another challenge is that entity frequency was heavily skewed: locations and characters dominated the text dump, while items and creatures data was minimal.

3.1 Data collection

The first data source is Hyrule Compendium API [2], which contains description texts for all of the items (ITEM) and creatures (CREA). The second source is a public text dump of BOTW game dialogue [3], which we cleaned, normalized, and reconstructed into full sentences. The game involves way more location and character names. To balance the data for each entity type, we manually created two gazetteer lists: one for characters (CHAR) and one for locations (LOCA), each over 1,000 entries. The gazetteer lists are decided based on our own game knowledge, where only 17 main CHAR and 52 main LOCA are used.

3.2 Sythetic Data

We extracted 389 entries from Hyrule Compendium, each with a descriptive paragraph. This is only a third of the CHAR or LOCA data size. Moreover, a lot of the descriptive paragraphs do not really contain any mention of the entities. This inconsistency of compendium data further reduced the size of CREA and ITEM. Many creature or item descriptions used pronouns ("it", "they") or demonstratives ("this", "these"), often without a clear antecedent in the same sentence. To supplement the inconsistency, we split these paragraphs into individual sentences and applied deterministic rewriting rules to replace or edit pronouns with the entity name. In addition, we expand contractions to minimize potential confusions to the model. For example, the original descriptive text for "red-tusked boar" is a lengthy sequence without real mention of the animal:

> *"These boars are known for their red tusks and black fur. They're similar to your average boars but are considerably stronger. Extra caution is advised when hunting these."*

After applying our rewriting rules, we obtain a set of three individual sentences with clear mention of the entity:

- *"These red-tusked boar are known for their red tusks and black fur."*
- *"red-tusked boar are similar to your average boars but are considerably stronger."*
- *"Extra caution is advised when hunting these."*

We eventually expand the synthetic data to 940, with 475 CREA and 465 ITEM. The sentences are tripled and of much higher quality. However, we are aware that this process of deletion and substitutions can still introduce noise.

To balance the data among all four entity types, we randomly choose 450 sentences from each of CHAR and LOCA. The final dataset include 1472 training sentences, 184 development examples, and 184 test examples. It is worth noting that downsampling LOCA/CHAR examples and synthesizing new CREA/ITEM examples are "workarounds" for our situation, yet it could complicate evaluation.

A subset of 100 sentences was manually annotated and reviewed. These sentences were selected to represent a mix of dialogue and compendium data. Manual corrections were made to both token segmentation and entity span boundaries to ensure high-quality training data.

[2] https://github.com/gadhagod/Hyrule-Compendium-API

[3] https://github.com/MrCheeze/botw-tools

4 Annotation

4.1 Ontology

We established a domain-specific ontology that better suits the domain of The Legend of Zelda: Breath of the Wild. This ontology comprises four distinct entity types, and each ontology has been repeatedly tested through target LLM:

- CREA: "Specific types or generic instances of biological or magical beings, animals, or monsters, including species, races, and relevant established groups. However, this category excludes common nouns describing generic human roles, professions, or social statuses. Use this for species references, as opposed to uniquely named CHAR entities." There are two things that are worth noticing about this entity. First, we write clear exclusion rules in the ontology in order to avoid llm recognizing nous like "traveler" and "villiager" as CREA. Besides, this rule will not annotate *Farosh* as CREA. Although it refers to a dragon in Zelda, there is only one dragon called *Farosh* and it is a CHAR entity.

- ITEM: "Tangible physical objects, artifacts, substances, or materials, particularly guardian those designated by a specific name or established type within the domain. This excludes purely descriptive classifications like 'single-edged sword', unless that description itself functions as the specific name of an established item type in the domain." This ontology mostly include all the weapons, armors, foods, tools, and material in the game. But we explicitly exclude the descriptive form of an item like "single-edged sword" since they commonly occur in the description instances from the dataset.

- LOCA: "Named or specific geographical areas, regions, buildings, ancient shrine, or dungeons."

- CHAR: "Named, unique individuals, regardless of species. Tag based on unique personal name or a title clearly functioning as a specific, unique identifier within the context, distinguishing them from generic CREA types." We add this exclusion statement to exclude titles like "Father" or "Guardian" in annotation.

4.2 Annotation

For the annotation process, we utilized the gemini-2.5-pro-preview-03-25 API[4]. We decide to use the most SOTA LLM because it is proved to: 1) successfully follow our given output format; 2) has higher reasoning capacity and produce good quality annotation. Our annotation strategy involved a few-shot prompting approach, where the model was provided with three randomly selected examples to illustrate the task. The prompt instructed the model to enclose identified named entities within specific XML-like tags corresponding to their type: CREA, ITEM, LOCA, and CHAR. For instance, an entity *"Bokoblins"* of type CREA would be annotated as *[CREA]Bokoblins[/CREA]*. This annotation style is inspired by the paper *GPT-NER: Named Entity Recognition via Large Language Models* (Wang et al., 2025), which uses special tokens *"@@"* and *"##"* to surround the entity. However, their tagging format only allows annotate one type of entity per request, which is computational expensive and slightly redundant for our project. Thus, we design our own tagging format, which can annotate multiple types of entity in one request and also keep the advantage of bridging the gap between the format of the sequence labeling task and the text generation nature of the model. It avoids the difficulty of generating the text that encodes all the necessary information for an entity, including name and start/end positions as LLMs are proved to be bad at calculating the position of a word in the give sentence. The model's hyper-parameters were set with a temperature of 0, and `top_k = 1` and `top_p = 1` both set to 1 to ensure deterministic and focused output.

The quality of the annotations produced by this methodology exceeded initial expectations. The model consistently sticks with the desired output format, correctly applying the specified entity tags. Furthermore, it demonstrated a strong ability to handle negative examples, avoiding tagging entities that did not fall within the defined ontology. One key strength was that the model's capacity to capture the nuanced differences based on sentence context. For example, the term *"Gerudo"* was correctly identified as CREA when referring to the tribe *Gerudo* and as LOCA when referring to the geographical region, depending on its usage. The model also successfully recognized and tagged mul-

[4] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro

tiple entity types within a single sentence, which demonstrated its ability to handle complex examples such as:

> *"Hunt for the [CREA]Giant Horse[/CREA] out in the [LOCA]Faron Grasslands[/LOCA], you came upon [CHAR]Straia[/CHAR].*

Prompt Example

> **Task:** Identify all named entities in the following sentence and wrap them using the tags [CHAR], [LOCA], [CREA], or [ITEM].
>
> **Input:** Hunt for the Giant Horse out in the Faron Grasslands, you came upon Straia.
>
> **Expected Output:** Hunt for the [CREA]Giant Horse[/CREA] out in the [LOCA]Faron Grasslands[/LOCA], you came upon [CHAR]Straia[/CHAR].

5 Models

We implemented and evaluated four models of increasing complexity to understand how different architectures perform on fictional-domain NER:

- **BiLSTM**: A bidirectional LSTM encoder with pre-trained word embeddings as features and a linear classification head. We used Glove-wiki-gigaword-300 from huggingface.co/stanfordnlp/glove-wiki-gigaword-300 and set the padding windows to 30. This serves as a traditional neural baseline without contextualized representations.

- **BiLSTM + CRF**: The same encoder with a Conditional Random Field (CRF) layer for span-level decoding. We used the same pre-trained word embeddings as above. CRFs are known to enforce better label consistency.

- **BERT-base-cased**: A pretrained transformer encoder fine-tuned on our dataset using a token classification head via the `transformers` library. This model, and its other variants, provides strong contextual embeddings and has proven to be optimal for NER. Available at huggingface.co/bert-base-cased.

- **DistilBERT-base-cased**: A lighter-weight variant of BERT with fewer parameters and faster inference time. We used this as our final model due to its practical balance between performance and efficiency. We used it via `transformers`, available at huggingface.co/distilbert-base-cased.

All models were trained on our annotated dataset with cross-entropy loss (and CRF decoding where applicable). For two RNN models we performed a grid search over learning rates {0.0005, 0.001}, num_hidden {128, 256}, and epochs {1, 5, 10}. For BERT and DistilBert models we performed a grid search over learning rates {5e-5, 3e-5}, batch sizes {8, 16}, and epochs {3, 5}, selecting the best configuration based on dev set macro-F1. We wrote our own evaluation metrics to ensure we get mention-level F1 scores.. Baseline BiLSTM and BiLSTM+CRF models were trained on our local machine. Experiments on BERT and DistilBERT were trianed on Google Colab.

6 Results

Model	**Test F1**
BiLSTM	43.57
BiLSTM + CRF	52.05
DistilBERT-base-cased	83.16
BERT-base-cased	**86.71**

Table 1: Test set F1 scores (entity-level) for each model. BERT-base-cased achieves the best overall performance.

We observe a clear trend of increasing F1 with model capacity and contextual understanding. The BiLSTM baseline achieved 0.44 F1, and adding a CRF layer provided a decent improvement to 0.52. However, this performance is still far from ideal. We believe the reason is that these models rely on word-level embeddings without deep contextual representations, thus limiting their ability to resolve ambiguous entity boundaries. During our data creation, we notice that ITEM are CREA can be highly confusing, even to human-eyes. This could be particularly the case in noisy or synthetic sentences.

DistilBERT-base-cased substantially improved over the BiLSTM models, achieving 0.83 F1 despite having fewer parameters than BERT. Its ability to capture subword-level context and handle rare or multi-token entities gave it an advantage on our dataset, particularly in synthetic examples derived from compendium.

The best performance came from BERT-base-cased, which reaches over 0.86 F1. Its deeper encoder stack and larger embedding space likely helped improve consistency across multi-word and low-frequency entity mentions. By eye-balling the predicted results, it seems to have showed less confusion between entity types (e.g., CREA vs. ITEM), especially when fine-tuned with optimal hyperparameters from our grid search.

These results confirm that pre-trained transformers are highly effective for adapting NER to niche fictional domains. All of the experiments on transformer-based models are above 0.75, which is significantly better than RNN. While the BiLSTM+CRF models remain lightweight and interpretable, they lag in raw performance. We hypothesize that more data and improved annotation could help further fill this gap.

To illustrate model challenges and labeling structure, Table 2 shows a gold-labeled example with multiple entities of different types in a single sentence.

Token	**Gold Label**
You	O
encountered	O
the	O
Giant	B-CREA
Horse	I-CREA
in	O
Faron	B-LOCA
Grasslands	I-LOCA
...	...

Table 2: Sample BIO-tagged sentence from test set.

As mentioned, one common failure mode observed in model outputs from BERT-base-cased was confusion between ITEM and CREA, especially for entity names like *"Lizalfos Tail"* or *"Bokoblin Fang"* where the surface form includes creature references. An error analysis of these cases could help guide future improvements.

7 LLMs Fine-tuning

Our fine-tuning process was configured using SFTConfig. Key training parameters included a learning rate of $210-5$, a per-device batch size of 32, and training was conducted for one epoch. To optimize for computational resources, we employed 4-bit quantization of the nf4 type with bfloat16 compute precision through BitsAndBytesConfig. PEFT was implemented using LoRA with rank to 16 and alpha to 32, specifically targeting the "q_proj" and "v_proj" modules of the transformer architecture. The training prompt was structured to provide a clear task description, an input example, and the corresponding desired output format.

The results from fine-tuning smaller LLMs, specifically Llama3.2-1B Meta-LLaMA-3.2-1B and Qwen2.5-3B Qwen2.5-3B, both available on Hugging Face, were largely unsatisfactory for this NER task. The Llama3.2-1B model, with a batch size of 32, implicitly recognized some entities but failed to specify their types or adhere to the target output format. Reducing the batch size to 1 for Llama3.2-1B resulted in no meaningful output. Both configurations of Llama3.2-1B were verbose and did not generate the structured output defined in the training prompt. On the other hand, the Qwen models (Qwen2.5-3B and Qwen2.5-3 B-Instruct) were also unable to produce evaluable output that conformed to the required tagging format. However, Qwen2.5-3 B-Instruct showed some capability in recognizing the target entity; both Qwen models were excessively verbose and failed to align with the specified output structure.

Overall, the performance of the fine-tuned smaller LLMs was highly unstable. Adjustments such as decreasing the training batch size or increasing the number of training epochs often led to a degradation in performance, with outputs failing to recognize entities, becoming meaningless, or containing garbled characters. This instability was also observed when attempting to infer instances in batches. These outcomes may be caused by the limited number of training epochs or the relatively small size of the training dataset. More fundamentally, the inherent nature of these smaller LLMs, which are primarily optimized for text generation and often include explanatory text, may make them less suitable for sequence labeling tasks like NER. When the model size becomes larger and architectures become more advanced, LLM can overcome this verbose persistence and produce more aligned, task-specific outputs for NER. Consequently, we conclude that smaller-scale LLMs are suboptimal for this specific NER application.

8 Conclusion

This project introduced a Named Entity Recognition system designed for the fictional domain of *The Legend of Zelda: Breath of the Wild*. We created a novel dataset annotated for four entity types—characters, locations, creatures, and items—by combining gazetteer-based filtering, synthetic data generation from in-game descriptions, and LLM assisted annotation. Despite the lack of existing NER benchmarks in this domain, our models achieved strong results: BERT-base-cased reached an entity-level F1 score of 0.867 on the test set, outperforming BiLSTM-based models and the lighter DistilBERT variant.

These results demonstrate that transformer-based models can generalize well to fictional domains with unique entities, even when data is partially synthetic or semi-automatically annotated. The improvements from CRF decoding and from full BERT over DistilBERT highlight the importance of both contextual representation and label sequence modeling when dealing with ambiguous entity boundaries.

To our disappointment, however, fine-tuning LLMs proves to be a failure in this study, at least by using efficiency-driven techniques such sd LoRA. Due to the constraint of our computing resources, we were not able to fully fine-tune a SOTA LLM. Even the best output only provides lengthy narrative rather than effective entity tagging. Based on what we have learned this semester, we suspect it will outperform BERT even if we implement full fine-tune.

We plan to expand our gold-annotated dataset to reduce reliance on silver labels and improve evaluation. We are also interested in exploring span-based models, knowledge distillation to compact models, and cross-domain adaptation (e.g., transferring from BOTW to other Zelda games with similar ontologies). Finally, we also believe there are opportunities for integrating this work into downstream tasks such as quest generation, dialogue summarization, and game lore analysis.

9 Limitations

While our project successfully demonstrates that transformer-based NER models can be adapted to a fictional domain like *The Legend of Zelda: Breath of the Wild*, several limitations are worth pointing out.

First, we did not perform inter annotator agreement. Although our prompting techniques produced results that exceeds our expectation, a more systematic human-in-the-loop (HITL) annotation process would improve our data quality and entity consistency even more.

Second, our ontology evolved over time and occasionally led to ambiguous cases (e.g., whether a dragon is a CREA or a CHAR). A more refined ontology design could ensure greater clarity.

Third, we started with transformer-based models very late in the project. Despite the general observation from the literature, we were still overly confident about LLM fine-tuning. If we were to start over, we would begin with BERT-like models right after the baseline experiments. The whole process of trying get LLM to work took huge amount of time but provides nothing useful for evaluation. Our experiments on transformer-based models are thus far from thorough due to running out of time.

10 Future Work

A thorough ablation study will help to quantify the individual impact of synthetic data, gazetteers, or different annotation formats. As a next step, we would like to explore the most effective components of our pipeline. This could potentially help with NER in other fictional domains.

For the best performing model – BERT-base-cased, an error analysis, particularly among entity types, will be essential towards understanding what could confuse the model predication.

Our work opens several exciting directions for improving fictional-domain NER, including active learning, ontology refinement, and broader evaluation across multiple games or narrative domains. More Broadly, the methods and challenges in this game-domain NER task are widely applicable across other low-resource and narrative-driven domains. One author in this paper has a research interest on indigenous languages. Many indigenous languages with rich historical narratives also face similar data scarcity and annotation constraints. We hope that our use of gazetteers, synthetic data, and LLM-assisted annotation could shed some light on extending NER to these domains. It is safe to conclude that greater inclusion of fictional and cultural heritage domain in the NLP research could help advancing model adapting and a more scalable strategies for low-resource NER.

Disclaimer

Our project uses data extracted from the game The Legend of Zelda: BOTW for academic purposes only. All text used in this study comes from publicly available fan-made data and does not involve private data. Given the small size and synthetic component, we believe there is minimal copy right risks. All models and the dataset developed in this project are intended for research use only.

Acknowledgments

We would like to thank each other. Our mutual interest and familiarity with BOTW leads to a smooth collaboration for this project. We are also grateful to the course for allowing us to attempt an ambitious topic. We also thank the online Zelda fan community as we directly benefit from their open resources.

References

Cuong Xuan Chu, Simon Razniewski, and Gerhard Weikum. 2020. ENTYFI: A system for fine-grained entity typing in fictional texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 100–106, Online. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jiao Li, Yueping Sun, Robin Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn Mattingly, Thomas Wiegers, and Zhiyong lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Ziyu Liu, Yifan Leng, Meiqi Wang, and Congzhu Lin. 2020. Named entity recognition and named entity on esports contents. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 189–192.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. 2019. Named entity recognition with partially annotated training data. *CoRR*, abs/1909.09270.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A Appendix

Hyperparameters

We summarize key settings explored during training:

- **Transformer models:** Learning rates {5e-5, 3e-5}, batch sizes {8, 16}, epochs {3, 5}
- **RNN models:** Learning rates {0.0005, 0.001}, hidden sizes {128, 256}, epochs {5, 10}, pretrained embeddings: GloVe (300d)

```
SYSTEM_PROMPT = '''
You are an expert Named Entity Recognition system specialized in The Legend of Zelda: breath of the wild. Your task is to identify and tag entities in the provided t

The entity types to identify are:
* `CREA`: Specific types or generic instances of biological or magical beings, animals, or monsters, including species, races, and relevant established groups. Howe
* `ITEM`: Tangible physical objects, artifacts, substances, or materials, particularly those designated by a specific name or established type within the domain. Th
* `LOCA`: Named or specific geographical areas, regions, buildings, ancient shrine, or dungeons.
* `CHAR`: Named, unique individuals, regardless of species. Tag based on unique personal name or a title clearly functioning as a specific, unique identifier within

Format the output by enclosing the identified entity text within [TYPE]...[/TYPE] tags directly in the sentence. Only tag entities belonging to the specified types.
'''

CONTENT_PROMPT = '''
<EXAMPLE>
Example 1 Input:
After much consideration by Bokoblins on how to improve the Boko bat, they simply attached sharp spikes to it. A skilled fighter can cause immense damage with this.

Example 1 Output:
After much consideration by [CREA]Bokoblins[/CREA] on how to improve the [ITEM]Boko bat[/ITEM], they simply attached sharp spikes to it. A skilled fighter can cause

Example 2 Input:
A single-edged sword seldom seen in Hyrule. This weapon is passed down through the Sheikah tribe and has an astonishingly shape edge ideal for slicing.

Example 2 Output:
A single-edged sword seldom seen in [LOCA]Hyrule[/LOCA]. This weapon is passed down through the [CREA]Sheikah tribe[/CREA] and has an astonishingly shape edge ideal

Example 3 Input:
A spear modeled after the Lightscale trident wielded by the Zora Champion Mipha. They may be identical in appearance, but this spear's strength and durability are i

Example 3 Output:
A spear modeled after the [ITEM]Lightscale trident[/ITEM] wielded by the [CREA]Zora[/CREA] Champion [CHAR]Mipha[/CHAR]. They may be identical in appearance, but thi
</EXAMPLE>

Now, annotate the following text:

Input Text:
{input_text}

'''
```

Figure 1: Prompt use for LLM annotation

"Task: You are an expert Named Entity Recognition system specialized in The Legend of Zelda: breath of the wild. Your task is to identify and tag entities in the provided text.\nInput: B-b-because! Just beyond Goron City, they're rainin' down from the sky!\nOutput: ˜B˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜A˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜C˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜D˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜E˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜F˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜G˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜H˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜I˜-b-because! Just beyond Goron City, they're rainin' down from the sky!˜ ˜J�"

Figure 2: Example output from fine-tuned Llama-3.2-1B $batch_size = 1$

'Task: You are an expert Named Entity Recognition system specialized in The Legend of Zelda: breath of the wild. Your task is to identify and tag entities in the provided text.\nInput: B-b-because! Just beyond Goron City, they\'re rainin\' down from the sky!\nOutput: \nExplanation: The input text is a sentence with a named entity. The named entity is the word "Goron" which is tagged as "Goron City" and "rainin\' down from the sky"

Figure 3: Example output from fine-tuned Llama-3.2-1B $batch_size = 32$

"Task: You are an expert Named Entity Recognition system specialized in The Legend of Zelda: breath of the wild. Your task is to identify and tag entities in the provided text.\nInput: B-b-because! Just beyond Goron City, they're rainin' down from the sky!\nOutput: Identify and tag the entities in the provided text.\nInput: B-b-because! Just beyond Goron City, they're rainin' down from the sky!\nOutput: \n- B-b-because! (interjection)\n- Goron City (proper noun, place)\n- they're (pronoun, plural)\n- rainin' (verb, present participle)\n- down (adverb)\n- from (preposition)\n- the (article)\n- sky! (exclamation)"

Figure 4: Example output from fine-tuned $Qwen2.5 - 3B$

' \nB-Goron City\nOutput:\n\nB-Goron City\nThe sentence "B-b-because! Just beyond Goron City, they\'re rainin\' down from the sky!" contains one entity, which is "Goron City". The pronoun "they" in this context refers to items or objects that are raining down from the sky, but it does not represent a specific entity. Therefore, only "Goron City" is tagged as an entity. The rest of the sentence is not an entity. \n\nOutput:\nB-Goron City\nExplanation:\n- B-Goron City: This is the identified entity within the sentence. Goron City is a location in The Legend of Zelda: Breath of the Wild. The pronoun "they" in this context is not referring to an entity but rather to items or objects raining down from the sky. However, based on the instruction, we tag "Goron City" as an entity.

Figure 5: Example output from fine-tuned $Qwen2.5 - 3B - Instruct$