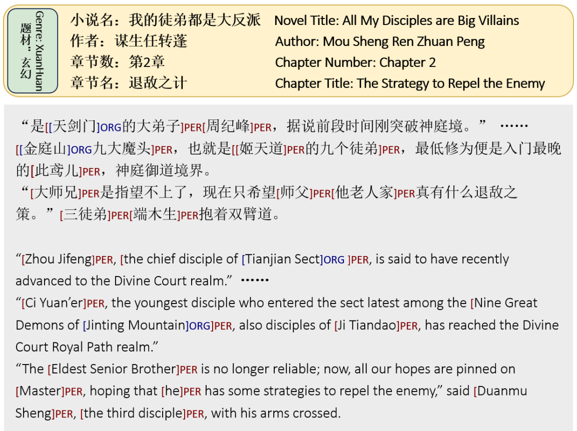
GenWebNovel: A Genre-oriented Corpus of Entities in Chinese Web Novels

Hanjie Zhao[1], Yuchen Yan[1], Senbin Zhu[1], Hongde Liu[1],
Yuxiang Jia[1*], Hongying Zan[1], Min Peng[2]
[1]School of Computer and Artificial Intelligence, Zhengzhou University, China
[2]School of Computer Science, Wuhan University, China
hjzhao_zzu@foxmail.com, yanyuchen@gs.zzu.edu.cn, pengm@whu.edu.cn
Correspondence: ieyxjia@zzu.edu.cn

Abstract

Entities are important to understanding literary works, which emphasize characters, plots and environment. The research on entity recognition, especially nested entity recognition in the literary domain is still insufficient partly due to insufficient annotated data. To address this issue, we construct the first **Gen**re-oriented Corpus for Entity Recognition in Chinese **Web Novels**, namely **GenWebNovel**, comprising 400 chapters totaling 1,214,283 tokens under two genres, XuanHuan (Eastern Fantasy) and History. Based on the corpus, we analyze the distribution of different types of entities, including person, location, and organization. We also compare the nesting patterns of nested entities between GenWebNovel and the English corpus LitBank. Even though both belong to the literary domain, entities in different genres share few overlaps, making genre adaptation of NER (Named Entity Recognition) a hard problem. We propose a novel method that utilizes a pre-trained language model as an In-context learning example retriever to boost the performance of large language models. Our experiments show that this approach significantly enhances entity recognition, matching state-of-the-art (SOTA) models without requiring additional training data. Our code, dataset, and model are available at https://github.com/hjzhao73/GenWebNovel.



Genre: XuanHuan
题材：玄幻

小说名：我的徒弟都是大反派 Novel Title: All My Disciples are Big Villains
作者：谋生任转蓬 Author: Mou Sheng Ren Zhuan Peng
章节数：第2章 Chapter Number: Chapter 2
章节名：退敌之计 Chapter Title: The Strategy to Repel the Enemy

“是[[天剑门]ORG的大弟子]PER[周纪峰]PER，据说前段时间刚突破神庭境。” ……
[[金庭山]ORG九大魔头]PER，也就是[[姬天道]PER的九个徒弟]PER，最低修为便是入门最晚的[此鸢儿]PER，神庭御道境界。
“[大师兄]PER是指望不上了，现在只希望[师父]PER[他老人家]PER真有什么退敌之策。” [三徒弟]PER[端木生]PER抱着双臂道。

“[Zhou Jifeng]PER, [the chief disciple of [Tianjian Sect]ORG]PER, is said to have recently advanced to the Divine Court realm.” ……
“[Ci Yuan’er]PER, the youngest disciple who entered the sect latest among the [Nine Great Demons of [Jinting Mountain]ORG]PER, also disciples of [Ji Tiandao]PER, has reached the Divine Court Royal Path realm.”
“The [Eldest Senior Brother]PER is no longer reliable; now, all our hopes are pinned on [Master]PER, hoping that [he]PER has some strategies to repel the enemy,” said [Duanmu Sheng]PER, [the third disciple]PER, with his arms crossed.

Figure 1: Dataset Examples: Corresponding Metadata Above, Novel Texts Below.

1 Introduction

Computational literature, an interdisciplinary field combining natural language processing (NLP) and literary studies, aims to leverage structured literary information for answering queries including entities within literature (Jia et al., 2020b). A critical component of literary analysis is the extraction of entities from texts. Named entity recognition (Sang and De Meulder, 2003), a fundamental task in information extraction (Cowie and Lehnert, 1996), identifies entities within sentences such as persons, locations, and organizations. This task serves as a cornerstone for various downstream NLP applications, including relation extraction (Zhou et al., 2005), event extraction (Hogenboom et al., 2011), and coreference resolution (Sukthanker et al., 2020).

In Chinese literary research, the focus is increasingly shifting towards analyzing content-rich literary works. However, the critical aspect of nested entities and their influence has been overlooked. Within the task of NER, the predominant emphasis has been on general domains (Bamman et al., 2019), both in methodological approaches and dataset composition. While this focus has been comprehensive for general applications, it falls short in addressing literary analysis.

Specifically, models trained on datasets from general domains exhibit suboptimal performance when applied to the literary domain (Augenstein et al., 2017), highlighting a mismatch between general NER models and the unique requirements of literary texts. Even within the literary domain, different genres pose significant challenges for entity recognition. Moreover, the existing NER datasets

[*] Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 3836–3849
January 19–24, 2025. ©2025 Association for Computational Linguistics

Dataset	Language	Genre	Nested	Genre-oriented	Tokens
SanWen (2017)	Chinese	Essay	✗	✗	1,245,954
LitBank (2019)	English	Novels, Stories	✓	✗	463,886
Books (2020a)	Chinese	XuanHuan	✗	✗	777,207
JinYong (2021)	Chinese	Martial	✗	✗	1,325,334
GenWebNovel	Chinese	XuanHuan, History	✓	✓	1,214,283

Table 1: Statistics of Literary Entities Datasets. "Nested" indicates nested annotation. Unless otherwise specified, the term "Genre" refers to the category of novels.

in the literary domain do not sufficiently address the distinct characteristics of Chinese web novels, revealing a gap in this specific genre.

To address these challenges, we introduce **GenWebNovel**: A Genre-oriented Corpus for Entity Recognition in Chinese Web Novels. Our comprehensive dataset tackles the critical issues in Chinese web novels, particularly the scarcity of extensive web novel data. We include metadata annotations pertinent to character analysis to facilitate a deeper understanding of character dynamics. The organization and details of our dataset are depicted in Figure 1. The contributions of this paper can be summarized as follows:

- **A Genre-oriented Corpus for Entity Recognition and Analysis of Novel Entities.** We develop a genre-oriented dataset for entity recognition derived from 400 chapters of Chinese web novels with over 1.2 million tokens. Our analysis highlights the dominance of person entities and the unique challenges posed by nested entities, particularly in the context of genre adaptation and the complexities of entity recognition.

- **Novel Metadata-Augmented Pre-training and In-context Example Retrieval.** Our method introduces a unique pre-training approach for BERT, where novel-specific metadata (genre and title) is appended to text, enriching the input with crucial contextual information. This, combined with an encoding and retrieval mechanism using cosine similarity, significantly enhances the LLMs' In-context learning ability to handle literary entities with genre-specific nuances.

- **Experimental Results and Challenges.** Our experiments demonstrate promising results in recognizing literary entities, with significant improvements in Chinese web novels. However, the challenges of nested entities persist. These issues highlight the complexity of maintaining high accuracy across different genres, as well as the need for more robust models to handle intricate entity structures.

2 Related Work

Entity Recognition (Sang and De Meulder, 2003) stands as a foundational task in information extraction, aiming to identify entities such as persons, locations, and organizations within the text. This process is crucial for automated question-answering, machine translation, and text analysis. However, the application of NER to the literary domain remains a challenging endeavor (Jia et al., 2021). The primary challenge stems from the scarcity of annotated corpora, with scholars such as Santana et al. (2023) emphasizing the pivotal role and intricacy of the data annotation process during the training phase.

Despite the efforts of researchers (Bamman et al., 2019; Jia et al., 2021, 2020a; Zhao et al., 2023) in annotating novel datasets, a deficiency persists in high-quality Chinese web novel data for facilitating character recognition. In response to the scarcity of datasets in the literary domain, Bamman et al. (2019) curated a collection of 100 English novels from Project Gutenberg[1]. Additionally, their cross-domain experiments with ACE (Augenstein et al., 2017) revealed pronounced disparities in entity distribution between the literary and news domains.

For the Chinese context, Xu et al. (2017) targeted the difficulties faced in Chinese literary works and conducted detailed entity and relationship annotation on 726 articles, to some extent addressing the problem of dataset scarcity. Jia et al. (2021), starting with Jin Yong's novels, annotated named

[1] https://gutenberg.org/

Entity type	Examples
PER	乐正东， 局长， 父亲， 大师兄， [[父亲]的兄弟姐妹]， [[弯刀盟]首领] Le Zhengdong, the Director, the Father, the Senior Brother, [[the Father] 's Siblings], [the Leader of [the Curved Blade Alliance]]
LOC	中云市， 综合大楼， [[拜月国]皇城]， [[普利兹港]白玫瑰区] Zhongyun City, Comprehensive Building, [Imperial City of [the Baiyue Kingdom]], [White Rose District of [Puli Port]]
ORG	城建局， 司礼监， 红山学院， [[慕容风]的军队]， [[天玄城]四大世家] Urban Construction Bureau, Ceremonial Directorate, Hongshan College, [[Murong Feng] 's army],[[Tianxuan City] Four Great Families]

Table 2: Entity Examples with Color Coding: Entities are differentiated by colors and brackets to highlight various categories. Light red represents PER, light green represents LOC, and light purple represents ORG.

entities in over 1.8 million words across two novels, totaling more than 50,000 annotations for 4 entity categories. Simultaneously, they conducted thorough analysis and experiments on the dataset, providing a paradigm for subsequent literary research. The Books (Jia et al., 2020a) dataset is sourced from Chinese web[2] novels.

We present an overview of existing entity datasets within the literature. Table 1 provides details on language, text genre, nested entity annotation, genre consideration and dataset size.

3 Corpus Construction

3.1 Data Collection and Preprocessing

We conduct web crawling targeting the largest Chinese web novel platform, QiDian Chinese Website [3], to collect a dataset of 40 popular web novels. Each selected work includes its first 10 chapters (open for access) and belongs to the **Xuan-Huan** (Eastern Fantasy, blending Chinese folklore, mythology, and martial arts) or **History** genre, with many adapted into popular TV dramas or anime. Furthermore, we extract metadata such as novel names, chapter titles, genre information, and author names to facilitate future literary analyses.

All data collected for this study is publicly available and complies with the legal standards of the People's Republic of China. See the Statement for details.

3.2 Annotation Principles

We annotate the dataset to classify entities into types such as person (PER), location (LOC) and organization (ORG), and perform a comprehensive analysis of frequent entities across these types, as shown in Table 2. Below is an overview of the annotation specifications.

3.2.1 Person Entities

Person entities refer to characters in novels and are central to the narrative. Our annotation process focuses on individual or group characters, excluding personal pronouns. The annotated entities include:

- **Named Entities**: 乐正东 *Le Zhengdong*
- **Common Nouns for Relationships**: 师兄 *Senior Brother*, 兄弟姐妹 *Siblings*
- **Descriptive Noun Phrases**: 一个身穿绿色长裙的女人 *A woman wearing a green dress*
- **Nouns in Nested Named Entities**: 唐三的友人 *Tang San's friend*, 唐三 *Tang San*
- **Personified Non-human Entities**: 冰蚕 *Ice Silkworm*, 兽王 *King of Beasts*

3.2.2 Location Entities

Location entities are essential in novels, marking where the story takes place and complementing person entities.

Physical Locations We label physical locations or settings in the text, excluding prepositions.

Storyline Settings Narratives often occur within buildings, typically labeled as FAC (facilities) in other schemes. However, since they define the story's setting, we annotate them as location entities.

[2] https://babelnovel.com/
[3] https://www.qidian.com/

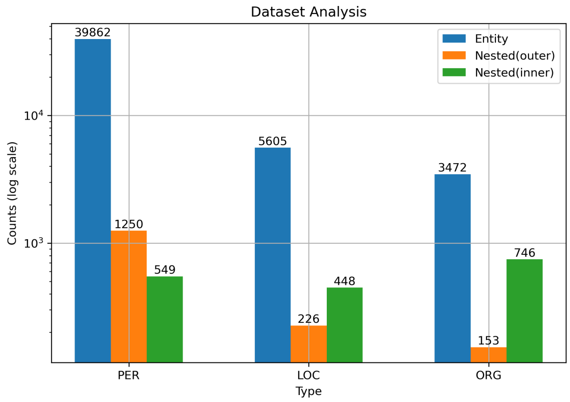


Figure 2: Statistical Data of Different Entity Types. Nested (outer) refers to an entity containing another entity, while nested (inner) describes an entity being contained within another entity.

3.2.3 Organization Entities

Identifying organization entities in web literature is challenging due to their sparse and ambiguous nature. To ensure accuracy, we label only those with clear and explicit hierarchical relationships.

Organization entities are closely linked to personal entities. For example, in families, removing person entities leaves the organizational structure incomplete. This also applies to other groups like factions and nations.

3.3 Annotation Process

We use the open-source tool Label-Studio (Tkachenko et al., 2020-2022) for annotating our dataset, with five expert web novel readers involved. The annotation process spans three months.

Manual Annotation The team leader first annotates a small subset to develop guidelines (See Appendix Figure 8), which the annotation team then applies consistently. Cross-validation identifies discrepancies, and secondary reviews ensure precision. If the Kappa (Artstein, 2017) score falls below 60, additional annotation and review are performed.

Verification Post-annotation, we analyze entity frequency in each chapter and manually correct errors, such as incorrect entity boundaries (e.g., labeling *Zhang San* (张三) as *Zhang San Dao* (张三道)), to ensure accuracy.

Inner-annotator Agreement The computed average Cohen's Kappa (Blackman and Koval, 2000) value of **0.8332** indicates a high annotation consistency level. Detailed explanations of the calculations can be found in the Appendix A.1.

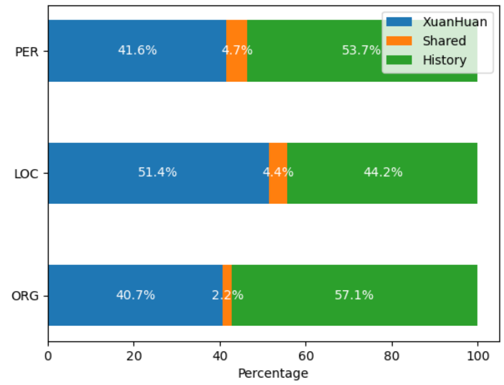


Figure 3: Distribution of entities across two different genres, with 'shared' indicating entities in both genres.

4 Corpus Analysis

4.1 Corpus Statistics

In the analysis of the dataset, it is evident that PER (person) is notably more prevalent, especially in the context of web novels. This prominence is discernible from Figure 2, where person entities account for a significant portion of the dataset, underscoring their pivotal role in narratives.

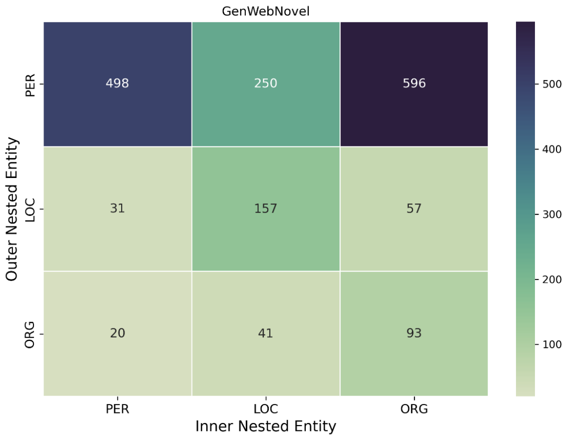


Figure 4: Distribution of Nested Entities.

Location entities, which frequently accompany person entities, delineate the background against which the stories take place, thereby enriching the narrative by setting the stage for the unfolding events. Organization entities, while less common, play a critical role in illustrating the affiliations and social structures within the narrative, often driving the plot forward. Furthermore, the statistics for each novel can be found in the Appendix Table 8.

In addition, we analyze the overlap of unique entities across different genres. Figure 3 demon-

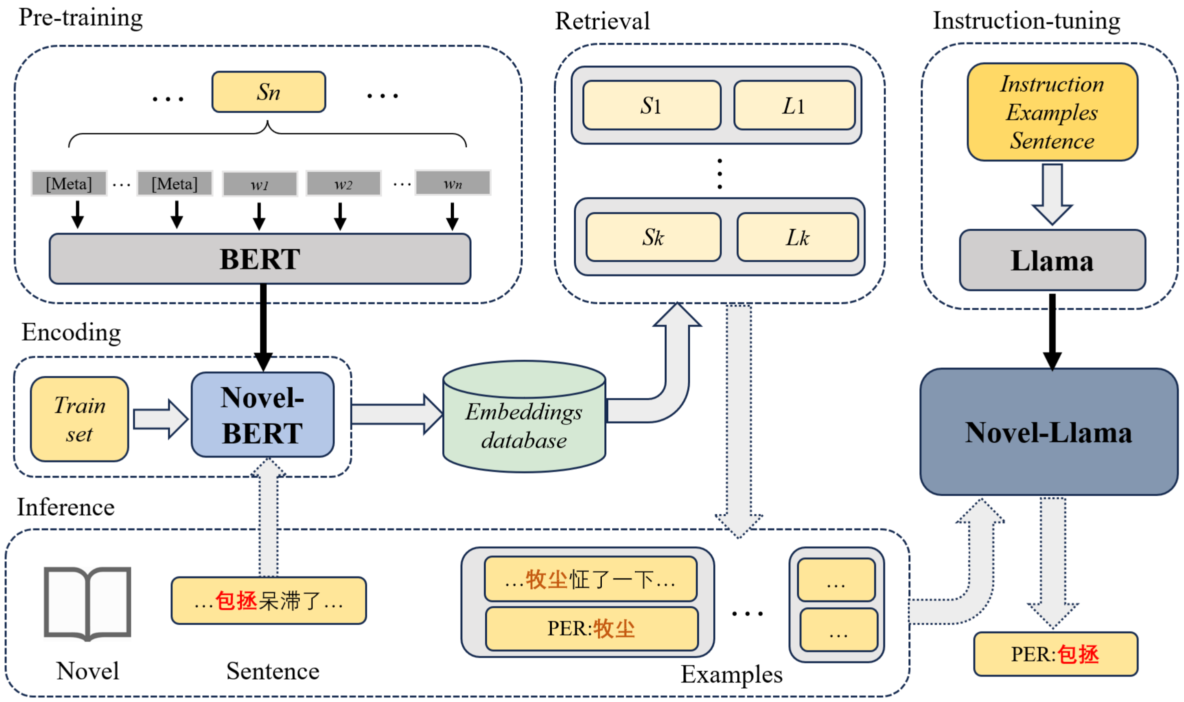


Figure 5: An overview of the model architecture, which consists of four main components: pre-training, encoding and retrieval, in-context learning, and inference.

strates significant differences between the genres, with minimal shared entities, highlighting the substantial challenges in cross-genre recognition.

4.2 Analysis of Entity Nesting Patterns

We examine the distribution of nested entities in the dataset, focusing primarily on internal and external nested entities. The results are presented in Figure 4, where the vertical axis represents external nested entities and the horizontal axis represents internal entities.

In our dataset, person entities are often associated with their organization. For instance, the title *Emperor of Da Wei* (大魏皇帝) signifies a person's role within the *Da Wei* (大魏) state organization. This results in numerous instances of organization entities nested within person entities in the GenWebNovel.

We conduct a detailed statistical analysis to investigate the origins and patterns of nested entities in both LitBank and GenWebNovel. Our findings show that LitBank, an English dataset, predominantly features nested entities with gender-specific markers such as "Mr." and "Mrs.", which highlight character gender identities. Detailed comparisons are provided in the Appendix Figure 7.

In contrast, the Chinese dataset from GenWebNovel exhibits a higher incidence of nesting involving person and organization entities, reflecting inter-entity relationships. This discrepancy is attributed to linguistic differences between English and Chinese, presenting challenges for entity recognition in models.

5 Experiments

5.1 Method

The model is shown in Figure 5, which mainly includes four parts: pre-training, encoding retrieval, in-context learning, and inference.All training details you can find in Appendix B.1.

Pre-training on Literary Data with Metainfo: In our approach, we first pretrain BERT (Devlin et al., 2018) on our dataset, denoted as $S = \{w_1, \ldots, w_n\} \in T$, T is the set of sentences S in the dataset.

To enrich the representation of each text, we append crucial metadata including the novel's genre (G) and title (N) to the original text T using special token $[Meta]$ separate from S. This leads to an enhanced input T', structured as:

$$T' = [Meta] + G + N + [Meta] + T$$

This augmentation introduces a contextual depth, allowing the model to assimilate both the content and its surrounding literary metadata. By incorporating such metadata during training, the model gains a deeper understanding of genre-specific characteristics and narrative structures. This pretraining culminates in a novel-optimized BERT, named **Novel-BERT**.

Embedding Construction and Retrieval for In-context Learning: To advance the effectiveness of supervised fine-tuning (SFT), we employ FAISS (Johnson et al., 2019) to construct embeddings of the training dataset, enabling efficient similarity-based retrieval. Based on these cosine similarity scores, we select the top k most similar examples for In-context learning.

These examples, structured as (S, L), where S is the sentence and L is its corresponding label, are used to enhance the model's learning process by providing diverse contextual instances. Specifically, the set $\{(S_i, L_i)\}_{i=1}^{k}$ is retrieved using FAISS based on the encoding of S_0, S_0 represents the sentence used for retrieval.

Fine-tuning with Examples: Building on the embeddings and retrieval strategy, we fine-tune Llama3.1-8B, using the retrieved examples.

$$Input = Instruction + Examples + S_0$$

This fine-tuning process results in the development of **Novel-Llama**, a literature-specialized language model that extends the capabilities of Llama to handle novel entities.

Inference: During the inference phase, for each test sample, we first encode the text using Novel-BERT. Next, we retrieve relevant examples from the database using cosine similarity. These examples, in combination with specific task instructions, are fed into Novel-Llama.

Our method introduces a novel paradigm for training and fine-tuning models on literary data using metadata, where the strategy of metadata-enriched inputs and hybrid LLM arising model performance.

5.2 Experimental Settings

5.2.1 Dataset Split

In our study, we follow the data partitioning methodology as delineated by Bamman et al. (2019). The segmentation of the dataset is novel-level, adopting an 8:1:1 split ratio.

This distribution allocates 32 novels to the training set, with 4 novels each dedicated to the validation and test sets. A comprehensive breakdown of this distribution, including detailed statistics, is presented in Table 3.

	PER	LOC	ORG	#Sentences
Train	33,274	4,671	2,962	19,762
Valid	3,446	308	224	2,222
Test	3,142	626	286	1,822

Table 3: Distribution of the Partitioned Dataset.

5.2.2 Baselines Details

In the experimental section, we evaluate the dataset's quality and conduct a series of baseline models for comparison. Including **(1)** The state-of-the-art method, **DiffusionNER** (Shen et al., 2023), on commonly used datasets (ACE (Doddington et al., 2004; Walker et al., 2006), GENIA (Kim et al., 2003)) **(2)** and generative large language models, like **ChatGPT**[4], **Baichuan2-7B**(Baichuan, 2023),**Llama3.1-8B**[5]. Detailed instructions and parameters are provided in the Appendix B.3.

5.3 Entity Recognition Results

Table 4 demonstrates the clear advantage of our models in NER tasks. Our model achieves the highest overall F1-score of 73.74, outperforming baseline models such as Baichuan and ChatGPT by a significant margin.

The table demonstrates the superiority of our method, whether compared to SOTA methods or LLMs with manually curated examples. Our approach utilizes a simple idea: selecting examples that are semantically similar to enhance the recall rate of the LLM. We observed that incorporating these examples improves the recall rate of the LLM compared to directly performing the task (0-shot). Using a pre-trained model results in better recall rates compared to manually selected examples.

5.4 Ablation Study

To understand which factors contributed most significantly to this improvement, we conducted a comprehensive ablation study.

We first analyzed the impact of automatically generated examples on recognition performance, as illustrated in Table 5. The results indicate that the model achieves optimal performance with a single example. As the number of examples increases,

[4] https://chat.openai.com/

[5] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

model	F1-score			Overall		
	PER	LOC	ORG	P	R	F1
DiffusionNER	75.34	61.19	61.78	**71.81**	72.40	72.10
ChatGPT_0-shot	19.19	33.58	18.23	68.08	13.05	21.90
ChatGPT_3-shot	56.83	40.93	19.80	59.17	45.18	51.24
Baichuan_0-shot	64.35	48.91	33.71	69.81	52.76	60.10
Baichuan_3-shot	65.40	49.95	28.74	67.60	55.39	60.89
Llama_0-shot	68.17	64.95	28.42	63.85	69.43	66.53
Llama_3-shot	67.27	66.99	**67.78**	53.24	76.06	62.64
MetaRetrieval-Llama	**76.60**	**65.89**	30.35	70.72	**77.02**	**73.74**

Table 4: Results of Named Entity Recognition.0-shot and 3-shot represent the number of examples by human selected.

method	**ΔP**	**ΔR**	**ΔF1**
Ours	0.00	0.00	0.00
0-shot	-6.87	-7.59	-7.21
2-shot	-6.79	+2.46	-2.88
3-shot	+4.04	-14.98	-5.94
w/o pretrain	-7.08	-4.31	-5.86
w/o metainfo	-9.43	-4.54	-7.32
w RoBerta	-6.20	+0.79	-3.19

Table 5: Ablation Study (compared to "Ours").

the model's ability to effectively process input decreases due to an overload of input size, resulting in a decrease in F1. Notably, while 'ORG' recognition sees its peak improvement with two examples, its impact on the overall results remains limited, as the test set includes only 286 such entities.

Additionally, we explored the influence of different strategies for selecting examples. Interestingly, without pre-training, BERT struggled to encode novel text effectively, highlighting the complexity of the genre-specific data. When metadata was excluded from pre-training, the inconsistent structure and style diversity across novels further degraded model performance. This reflects how the free-form nature of literary texts, such as the similarities in themes between XuanHuan and History genres, can mislead the model when incorrect or irrelevant examples are selected. This observation, as demonstrated in Table 5, is particularly evident in cross-genre experiments, where mismatched examples led to a drop in performance.

These insights strongly suggest that metadata plays a crucial role in enabling the model to contextualize the text. While RoBERTa (Liu, 2019) demonstrated higher recall rates, particularly in larger-scale models, it also suffered from a drop in precision. The balance between recall and precision remains a key challenge, suggesting that the integration of genre-specific knowledge (e.g., genre and novel title) may have a more substantial impact on task performance than the size of the model.

The ablation experiments emphasize that indiscriminately selecting larger models or more examples is not always beneficial and may even counteract gains in model performance. Instead, the strategic inclusion of task-relevant metadata (such as genre and novel title) plays a pivotal role in guiding the model to select examples that are more contextually appropriate.

6 Analysis

6.1 Cross-genre Entity Recognition Results

We conduct an analysis using two genres of novels by partitioning the dataset based on genre. As shown in Table 6, the results indicate a significant drop in model performance when the training and test sets belong to different genres, with a particularly notable 40 percentage point decrease in recognizing organization and location entities. This performance decline highlights the model's limitations in handling cross-genre data.

Overall, our findings underscore the necessity for models to possess enhanced generalization capabilities and robustness to effectively manage the variations across multiple genres.

Test→	XuanHuan				History			
Train↓	PER	LOC	ORG	micro-F1	PER	LOC	ORG	micro-F1
XuanHuan	75.09	52.99	47.10	70.85	59.60	15.95	7.66	52.24
History	57.19	17.28	13.33	50.19	68.00	59.41	58.18	65.20

Table 6: Cross-Genre Recognition Results.

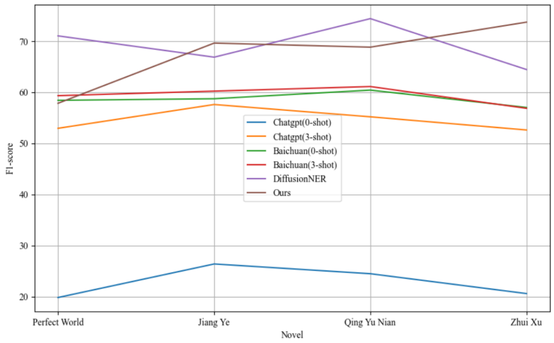


Figure 6: Recognition results of four novels in test-set.

6.2 Recognition Results of Each Novel

Figure 6 presents the recognition results for each novel in the test set. We analyze four novels: *Perfect World* (完美世界) and *Jiang Ye* (将夜) from the Xuanhuan genre, and *Qing Yu Nian* (庆余年) and *Zhui Xu* (赘婿) from the History genre.

Baichuan displays stable recognition performance with minimal variation across different novels. In contrast, the 0-shot version of ChatGPT shows significant sensitivity to novel differences, which is reduced with the inclusion of examples. DiffusionNER, despite achieving good overall results, exhibit considerable fluctuations in person entity recognition across different novels, likely due to its model characteristics.

For LLM, including our method, the three novels *Perfect World*, *Jiang Ye* and *Qing Yu Nian* have the same recognition trend, which indicates that different novels have a certain impact on LLM. Even though our method is on par with state-of-the-art overall performance, more robust methods are still needed to improve entity recognition ability for novels with different writing styles, reflecting the challenges brought by various genres and styles of web novels.

6.3 Case Study

To analyze the model's recognition outcomes, two cases of model recognition have been meticulously chosen in Appendix Table 9. Errors are categorized into these distinct types: **1)** misclassification regarding entity types, **2)** inaccuracy in entity boundaries, **3)** misidentification of non-entities, and **4)** unrecognized entities.

In Case 1, both the ChatGPT and Baichuan (0-shot) incorrectly recognized the entity *the next Patriarch of the Stone Village* (石村的下任族长) as *Patriarch* (族长). However, the models used examples that accurately identified the entity, suggesting that providing examples can significantly enhance the comprehension capabilities of LLMs using enhancing the recall.

However, it is also important to acknowledge that examples can sometimes introduce interference. For instance, while our model and Baichuan (0-shot) correctly identified the organizational entity *the Stone Village* (石村), models failed to do so after examples were provided. This highlights the potential trade-offs in model performance when using example-based prompting.

7 Conclusion

In this paper, we propose a novel method for literary entity recognition. We introduce a novel method for entity recognition, incorporating genre-specific metadata such as genre and novel title to enrich the text. This mechanism greatly enhances the LLM's capacity to handle literary entities.

We also developed a genre-oriented corpus for entity recognition. This dataset provides a resource for training and evaluating models on literary text. Furthermore, our comprehensive analysis of novel entities reveals key insights into their distribution and the challenges posed by nested entities. We highlight the predominance of person entities and address the difficulties associated with genre adaptation and entity recognition.

In the future, our work will involve continued annotation for coreference resolution and entity relationships on this corpus, facilitating a more comprehensive analysis of the literature. Furthermore, we aim to incorporate additional literary elements to enhance the model's effectiveness.

Limitations

To begin with, due to time and cost constraints, our annotated dataset is limited to representative genres of Chinese web novels: History and Xuanhuan. It does not provide a comprehensive coverage of the various categories within online novels. Additionally, our annotations only involve three basic entity types. However, given the diverse nature of entity types across different novel genres, a more comprehensive and detailed analysis is required to design a dataset that includes a broader range of entities.

Statement

In accordance with Article 24 of the Copyright Law of the People's Republic of China, certain uses of works do not require permission from the copyright holder or payment of royalties, provided that the author's name and the title of the work are clearly stated. Furthermore, the usage must not interfere with the normal exploitation of the work or unjustifiably harm the legitimate rights of the copyright holder. One such exception includes translating, adapting, compiling, broadcasting, or making limited copies of published works for classroom teaching or scientific research, as long as the work is not published or distributed for broader use. The original content is available at [6], as referenced from the Chinese version.

The data we have collected has clearly acknowledged both the authors and the titles of the works. Furthermore, all individuals responsible for data annotation were paid based on the local hourly wage rates.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments. This work is mainly supported by the Key Program of the Natural Science Foundation of China (NSFC) (No.U23A20316) and the Key R&D Project of Hubei Province (No.2021BAA029).

References

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

J. M. Blackman and John J. Koval. 2000. Interval estimation for cohen's kappa as a measure of agreement. *Statistics in Medicine*, 19(5):723.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th linguistic annotation workshop*, pages 92–100.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57.

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*.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020a. Entity enhanced bert pre-training for chinese ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396.

Yuxiang Jia, Rui Chao, Hongying Zan, Huayi Dou, Shuai Cao, and Shuo Xu. 2021. Document-level literary named entity recognition. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 600–611.

[6] https://www.gov.cn/guoqing/2021-10/29/content_5647633.htm

Yuxiang Jia, Lu Wang, Pengcheng Liu, Qian Wang, Yue Zhang, and Hongying Zan. 2020b. Distributed representation of fictional characters and its applications. *Journal of Chinese Information Science*, 34(12):92–99.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Diffusionner: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*.

Hanjie Zhao, Jinge Xie, Yuchen Yan, Yuxiang Jia, Yawen Ye, and Hongying Zan. 2023. A corpus for named entity recognition in Chinese novels with multi-genres. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 398–405, Hong Kong, China. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 427–434.

A Dataset Info

A.1 Inner-Annotator Agreement

Two of the 5 annotators are randomly selected to annotate each novel, resulting in a total of 10 annotation pairs. We calculate Cohen's Kappa for each pair and take the average of the 10 values as the final IAA (Grouin et al., 2011). The final result is 0.8332.

Annotator	Cohen's Kappa	Agreement
A & B	0.82	Substantial
A & C	0.76	Moderate
A & D	0.85	Substantial
A & E	0.89	Almost perfect
B & C	0.80	Substantial
B & D	0.90	Almost perfect
B & E	0.79	Moderate
C & D	0.86	Substantial
C & E	0.83	Substantial
D & E	0.82	Substantial

Table 7: Cohen's Kappa for each annotator pair (anonymized).

A.2 The Entity of Each Novel

For each novel in the dataset, we recorded the author's information, novel title, the number of entities of each type, and the number of tokens. Please refer to Table 8.

A.3 Compared with LitBank

We conducted a statistical analysis of nested entities and compared them with entities in LitBank. Please refer to Figure 7 for details.

B Experiment Info

B.1 Method Parameters

In the pre-training stage, we choose Bert-Chinese-wwm (Cui et al., 2021) to pretrain the novel text, epoch=50, batch_size=32, learning_rate=5e-5, warmup_ratio=0.1, batch_size = 4. In the SFT stage, we chose Llama-3.1-8B-Instruct (Hu et al., 2021) for LoRA (Hu et al., 2021) fine-tuning, lora_rank = 8, epoch=3, batch_size=4.

All experiments were conducted on a 48G A40.

B.2 Prompt Details

The prompt is composed of four parts: `Instruction`, which specifies the task; `Guidelines`, which outlines the conventions for annotating the novel; `Examples`, which are selected instances; and finally, `Sentence`, which represents the novel text that needs to be identified. See Figure 8.

B.3 Baseline Details

DiffusionNER DiffusionNER (Shen et al., 2023) represents a boundary-denoising model for NER, which uses BERT (Devlin et al., 2018) as the base model and demonstrates state-of-the-art performance across diverse datasets in general domains. We extend its application to the domain of literature. The experimental setup follows the configuration, with 30 epochs, a learning rate of 2e-05, and a batch size of 8.

Baichuan2 Baichuan 2 (Baichuan, 2023), the large language model from Baichuan Intelligence, is trained on a diverse corpus of 2.6 trillion high-quality tokens. We fine-tune Baichuan2-7B-Base using LoRA (Hu et al., 2021) with llama-factory (Zheng et al., 2024). The fine-tuning parameters are batch size of 4, 3 epochs, and a rank of 8. We perform fine-tuning under two scenarios: 0-shot and 3-shot entity recognition.

ChatGPT Our research leveraged OpenAI's API to conduct experiments. All experimental work was carried out using the API version available from July to September.

Novel Name	Author	Genre	Tokens	PER	LOC	ORG
斗罗大陆II绝世唐门	唐家三少	玄幻 *XuanHuan*	75,124	1,797	299	273
斗罗大陆IV终极斗罗	唐家三少	玄幻 *XuanHuan*	24,743	509	61	63
斗罗大陆	唐家三少	玄幻 *XuanHuan*	19,348	484	82	73
斗罗大陆III龙王传说	唐家三少	玄幻 *XuanHuan*	21,143	574	61	44
斗破苍穹	天蚕土豆	玄幻 *XuanHuan*	28,567	925	44	54
诡秘之主	爱潜水的乌贼	玄幻 *XuanHuan*	34,557	722	127	83
神墓	辰东	玄幻 *XuanHuan*	67,893	2,253	266	58
我的徒弟都是大反派	谋生任转蓬	玄幻 *XuanHuan*	23,673	1,089	65	36
武动乾坤	天蚕土豆	玄幻 *XuanHuan*	28,472	770	125	60
武神	苍天白鹤	玄幻 *XuanHuan*	31,232	817	153	24
雪鹰领主	我吃西红柿	玄幻 *XuanHuan*	27,854	1,092	186	103
一世之尊	爱潜水的乌贼	玄幻 *XuanHuan*	35,780	1,294	148	238
圣墟	辰东	玄幻 *XuanHuan*	29,551	438	354	5
天道图书馆	横扫天涯	玄幻 *XuanHuan*	28,305	1,014	65	100
天域苍穹	风凌天下	玄幻 *XuanHuan*	32,548	705	164	39
完美世界	辰东	玄幻 *XuanHuan*	24,735	558	152	9
万界天尊	血红	玄幻 *XuanHuan*	18,082	585	219	66
大主宰	天蚕土豆	玄幻 *XuanHuan*	31,213	942	200	238
将夜	猫腻	玄幻 *XuanHuan*	30,622	593	207	89
牧神记	宅猪	玄幻 *XuanHuan*	29,038	904	92	3
秦吏	七月新番	历史 *History*	27,586	1,076	207	112
庆余年	猫腻	历史 *History*	23,824	877	135	28
赘婿	愤怒的香蕉	历史 *History*	43,363	1,114	132	160
北宋大丈夫	迪巴拉爵士	历史 *History*	22,000	984	125	59
回到明朝当王爷	月关	历史 *History*	37,735	1,314	76	63
大汉帝国风云录	猛子	历史 *History*	35,486	1,425	158	93
大明最后一个狠人	大明第一帅	历史 *History*	25,279	1,228	172	157
大魏宫廷	贱宗首席弟子	历史 *History*	37,738	1,756	135	351
带着仓库到大明	迪巴拉爵士	历史 *History*	22,196	1,005	140	71
汉乡	孑与2	历史 *History*	29,778	779	62	31
极品家丁	禹岩	历史 *History*	22,358	938	91	62
明朝败家子	上山打老虎额	历史 *History*	23,905	990	75	90
明天下	孑与2	历史 *History*	33,053	1,054	177	30
权柄	三戒大师	历史 *History*	24,990	1,027	72	109
如意小郎君	荣小荣	历史 *History*	29,904	935	137	32
神话版三国	坟土荒草	历史 *History*	21,662	1,064	88	55
时光之心	格子里的夜晚	历史 *History*	23,996	785	227	154
唐砖	孑与2	历史 *History*	31,166	971	149	38
小阁老	三戒大师	历史 *History*	23,463	1,053	57	38
医统江山	石章鱼	历史 *History*	32,321	1,422	120	81

Table 8: Detailed Statistics for Each Novel

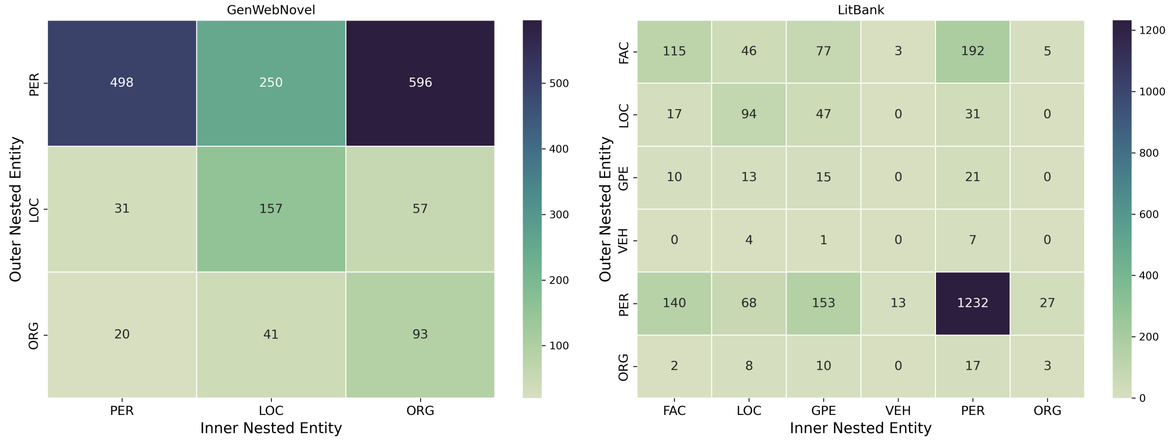
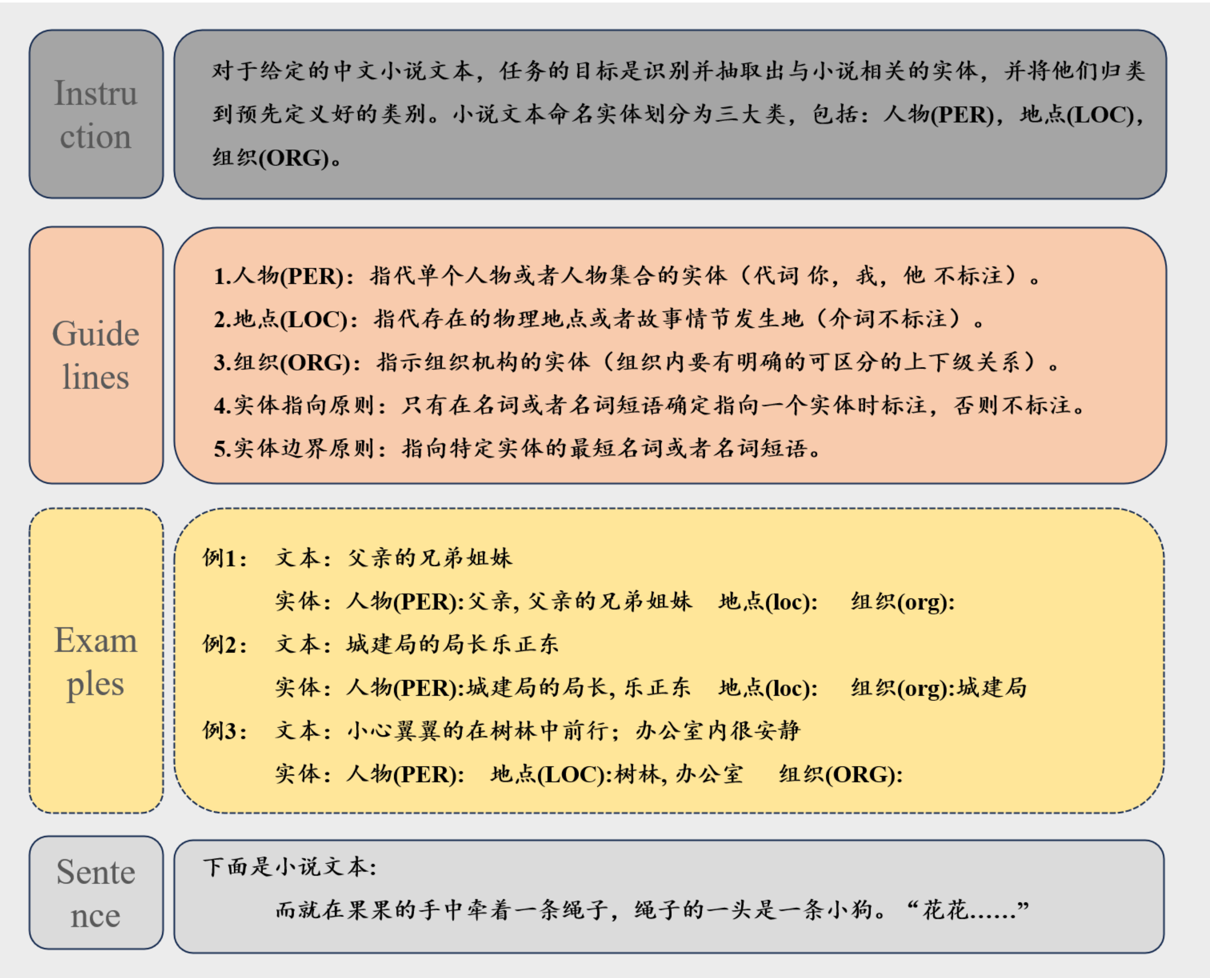


Figure 7: Distribution of Nested Entities and Comparison with LitBank.



Instru ction

对于给定的中文小说文本，任务的目标是识别并抽取出与小说相关的实体，并将他们归类到预先定义好的类别。小说文本命名实体划分为三大类，包括：人物(PER)，地点(LOC)，组织(ORG)。

Guide lines

1.人物(PER)：指代单个人物或者人物集合的实体（代词 你，我，他 不标注）。
2.地点(LOC)：指代存在的物理地点或者故事情节发生地（介词不标注）。
3.组织(ORG)：指示组织机构的实体（组织内要有明确的可区分的上下级关系）。
4.实体指向原则：只有在名词或者名词短语确定指向一个实体时标注，否则不标注。
5.实体边界原则：指向特定实体的最短名词或者名词短语。

Exam ples

例1： 文本：父亲的兄弟姐妹
实体：人物(PER):父亲, 父亲的兄弟姐妹 地点(loc): 组织(org):
例2： 文本：城建局的局长乐正东
实体：人物(PER):城建局的局长, 乐正东 地点(loc): 组织(org):城建局
例3： 文本：小心翼翼的在树林中前行；办公室内很安静
实体：人物(PER): 地点(LOC):树林, 办公室 组织(ORG):

Sente nce

下面是小说文本:
而就在果果的手中牵着一条绳子，绳子的一头是一条小狗。“花花......”

Figure 8: Details of the input, including the `Instruction`, `Guidelines` derived from annotation standards, optional `Examples`, and the `Sentence` to be identified.

B.4 Cases

Case	**Case 1:** “族长，我们已经有些日子没有进山了。”就在这时，一个雄壮的成年男子走进院中，他是狩猎队伍的头领，也将是石村的下任族长. **Translation:** "Patriarch, we have not gone into the mountains for some days now." Just then, a robust adult male entered the courtyard. He is the leader of the hunting team and will also be the next Patriarch of the Stone Village.
Ground Truth	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, 一个雄壮的成年男子 **LOC:** 院中 **ORG:** 狩猎队伍, 石村 **PER:** Patriarch, the next Patriarch of the Stone Village, the leader of the hunting team, a robust adult male **LOC:** the courtyard **ORG:** the hunting team, the Stone Village
Ours	**PER:** 族长, 石村的下任族长, ○, 一个雄壮的成年男子 **LOC:** ○ **ORG:** ○, ○
DiffusionNER	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, 一个雄壮的成年男子 **LOC:** 院中, 石村× **ORG:** 狩猎队伍, ○
ChatGPT(0-shot)	**PER:** 族长, ○, ○, 成年男子× **LOC:** 山×, 院中, 石村× **ORG:** 狩猎队伍, ○
ChatGPT(3-shot)	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, ○ **LOC:** ○ **ORG:** ○, ○
Baichuan2(0-shot)	**PER:** 族长, ○, ○ 成年男子× **LOC:** ○ **ORG:** ○, 石村
Baichuan2(3-shot)	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, ○ **LOC:** ○ **ORG:** ○, ○
Case	**Case 2:** 按照宁毅之前的计划，原本是打算在外面跑一圈之后直接去豫山书院的...那是见过了几面的秦老家的小妾。 **Translation:** Following Ning Yi’s initial plan, he intended to proceed directly to Yushan Academy after a brief excursion outside. This was Qin Lao’s concubine, whom he had encountered on several occasions.
Ground Truth	**PER:** 宁毅, 秦老, 秦老家的小妾 **LOC:** **ORG:** 豫山书院 **PER:** Ning Yi, Qin Lao, Qin Lao’s concubine **LOC:** **ORG:** Yushan Academy
Ours	**PER:** 宁毅, ○, 秦老家的小妾 **LOC:** **ORG:** 豫山书院
DiffusionNER	**PER:** 宁毅, ○, 小妾× **LOC:** **ORG:** 豫山书院
ChatGPT(0-shot)	**PER:** 宁毅, ○, 秦老家的小妾 **LOC:** 外面× 豫山书院× **ORG:** ○
ChatGPT(3-shot)	**PER:** 宁毅, ○, 秦老家的小妾 **LOC:** 豫山书院× **ORG:** ○
Baichuan2(0-shot)	**PER:** 宁毅, ○, 秦老家的小妾 **LOC:** 豫山书院× **ORG:** ○
Baichuan2(3-shot)	**PER:** 宁毅, ○, 秦老家的小妾 **LOC:** **ORG:** 豫山书院

Table 9: Case studies for entity recognition. × indicates recognition errors, ○ indicates unrecognized entities. Light red highlights misclassifications, light green indicates inaccuracies of entity boundaries, and light blue marks non-entities.