MythNER: Multi-Agent NER Extraction and Benchmarking in Chinese Myth Narratives

Anonymous ACL submission

Abstract

Named entity recognition (NER) performs strongly in well-studied domains, yet mythological narratives pose a long-tail setting where entityhood is defined by referable instances, mentions vary through titles and aliases, and correct extraction requires stable long-span boundaries. We introduce MythNER, a culturally grounded NER benchmark for Chinese mythology built from *Ne Zha* subtitle text. MythNER uses four flat labels (PER/LOC/ORG/OBJ) with conservative annotation criteria; in particular, OBJ is restricted to *named* artifacts and fixed technique titles rather than generic concepts. We evaluate a zero-shot Chinese spaCy model, a supervised BERT token-classification baseline, and a multi-agent LLM extraction pipeline with chunk/context tuning on the held-out *Ne Zha Part 2* test set. Results show substantial domain shift for off-the-shelf NER, strong gains from supervised adaptation, and further improvements from agentic extraction under well-chosen constraints. Our analysis characterizes dominant failure modes—boundary drift under exact-span scoring, mythology-specific type ambiguity, and over-extraction of generic nouns as OBJ—highlighting the need for iterative consistency checks in narrative-domain NER.

1 Introduction

Named entity recognition (NER) is a foundational component for information extraction, question answering, and knowledge base construction (Li et al., 2022; Jehangir et al., 2023).

While modern NER systems perform strongly in well-studied domains such as newswire, their assumptions often break in narrative text. Narratives routinely exhibit aliases and titles, implicit references, invented lexicons, and long-range dependencies in which the identity and type of an entity is only resolved after additional context (Bamman et al., 2019; Chu et al., 2020).

In this regime, accurate tagging is less a single-pass string labeling problem and more an exercise in *iterative consistency checks*: predictions must remain coherent across repeated mentions, shifts in viewpoint, and dispersed evidence. This motivates workflows that go beyond local extraction, pairing chunk-level recognition with document-level consolidation and verification to enforce global consistency.

Chinese mythology provides a particularly revealing stress test for narrative NER. Mythological texts, for instance fantasy movie subtitles, contain culturally grounded naming conventions, honorific- and role-heavy mentions, and named artifacts or techniques that behave like proper names, all of which can blur boundaries and types. For Chinese, these challenges are compounded by segmentation ambiguity and compact mentions, increasing the likelihood of span and label confusions (Gui et al., 2019).

At the same time, existing narrative NER resources tend to emphasize either Western myth and fantasy (similar narrative structure but different cultural and linguistic grounding) or Chinese fictional literature (shared language but a different entity ecology), leaving a gap for Chinese mythology. To address this gap, we introduce MythNER, a benchmark built from Chinese subtitles of *Ne Zha* films (Parts 1–2), designed to evaluate myth-domain NER under a controlled narrative setting.

We benchmark off-the-shelf zero-shot NER, a supervised encoder baseline, and an agentic extraction pipeline that decomposes long-context NER into chunked extraction, global

consolidation, and verification/correction (Wang et al., 2025b; Zhang et al., 2024). Across this ladder, off-the-shelf tools struggle under myth-domain shift, supervised training improves robustness, and agentic extraction yields further gains by enforcing cross-chunk consistency.

Our contributions are:

- **Dataset:** a culturally grounded Chinese mythology NER benchmark with clear guidelines and a coarse schema (PER/LOC/ORG/OBJ).
- **Benchmark evidence:** a controlled comparison that characterizes the gap between off-the-shelf NER and in-domain modeling for myth narratives.
- **Method:** an agentic extraction pipeline that performs chunked extraction followed by global consolidation and verification/correction to improve document-level consistency.

2 Related Work

Named entity recognition has progressed from lexicon- and rule-based systems to statistical sequence labeling, neural encoders, and transformer-based representations (Li et al., 2022; Jehangir et al., 2023; Pakhale, 2023). Despite strong performance on mainstream benchmarks, fictional and mythological narratives remain challenging: entity identity is shaped by evolving roles, aliases, and world-specific ontologies, and errors compound when the same character must be tracked consistently across long contexts. Literary and fiction-oriented resources highlight domain-specific entity definitions and distributions (Bamman et al., 2019), and practical systems for fiction emphasize the need for domain-specific typing and consolidation strategies (Chu et al., 2020). Complementary analyses in other narrative-like domains (e.g., tabletop role-playing corpora) likewise show systematic differences from standard NER settings (Weerasundara and de Silva, 2023), and fine-tuning studies in fantasy settings further support the value of in-domain adaptation (Sivaganeshan and De Silva, 2023).

For Chinese and historically oriented text, persistent issues include segmentation ambiguity and sparse supervision; lexicon-aware modeling remains a common mitigation for Chinese NER (Gui et al., 2019). Adjacent evaluations in ancient Chinese NER further illustrate both the progress and the domain gaps that arise once genre, period, and entity taxonomy shift (Li et al., 2025). Beyond modeling, annotation studies and multi-genre benchmarks underline that entity definitions are not purely technical: label sets and annotator interpretation can vary, affecting both training and evaluation (Tedeschi and Navigli, 2022; Peng et al., 2024).

Recent work explores using large language models (LLMs) directly for NER in zero-/few-shot settings (Wang et al., 2025a). However, narrative-scale extraction stresses long-context tracking and cross-mention consistency, motivating structured agentic alternatives. Multi-agent frameworks decompose extraction into specialized roles (e.g., proposal, verification against an ontology, aggregation), which is especially relevant when outputs must satisfy schema constraints (Tao et al., 2025; Wang et al., 2025b). Long-context collaboration mechanisms provide another path to maintaining state across long documents (Zhang et al., 2024; Zhuang et al., 2025).

3 Dataset and Annotation

3.1 Corpus

MythNER is a culturally-grounded named entity recognition benchmark built from Chinese subtitle text from the animated *Ne Zha* films. The domain is a Chinese mythological narrative world (e.g., *Fengshen*-related traditions and *Journey to the West*-related motifs), which is widely represented across Chinese myth literature and popular media. We use this setting as a representative high-context domain in which entity mentions include characters (PER), places (LOC), organizations/factions (ORG), and named artifacts or technique titles (OBJ).

3.2 Task and label schema

We perform coarse-grained NER with four flat labels: PER, LOC, ORG, and OBJ. The task follows standard span-based NER: given a subtitle sequence, the model predicts entity spans and assigns one of the four types.

3.3 Data split

We split the corpus by film to evaluate generalization across sequels. Specifically, we use *Ne Zha Part 1* (P1) for training and validation, and *Ne Zha Part 2* (P2) as a held-out test set. Within P1, we randomly split into 80% training and 20% validation. The P1 split is used for training and tuning supervised baselines. The entire P2 test set is held out for evaluation, including fair comparisons for zero-shot spaCy and our LLM-agent methods.

3.4 Dataset statistics and entity overlap

Figure 1 summarizes the dataset size and label distribution. Overall, MythNER contains 697 entity mentions (159 unique entities) under the four-label schema. The held-out P2 test set contains 415 mentions (106 unique), with per-label totals of 150 PER, 90 ORG, 75 LOC, and 100 OBJ.

To quantify cross-film generalization difficulty, we measure unique-entity overlap between training and test. Approximately 31.1% of unique entities in the test set also appear in training, implying that most test-set unique entities are unseen during training.

3.5 Annotation workflow

MythNER is annotated by three annotators and one adjudicating judge. Annotation is conducted in Doccano, self-hosted on a Google Cloud VM. Annotators label entity spans and coarse types (PER/LOC/ORG/OBJ). Ambiguous cases are resolved via discussion to align on guidelines, after which the judge performs final review and consolidation.

3.6 Annotation guidelines

We follow a flat NER schema without nested or overlapping spans. We adopt conservative criteria for entityhood: mentions should refer to an independently identifiable instance in context. We exclude purely event/phenomenon expressions, and treat OBJ as a strict category that covers named artifacts and fixed technique titles rather than generic concepts. When a surface form of a location word can plausibly be interpreted as a place or an institution depending on its syntactic role, we consistently annotate it as LOC. When a full canonical name appears, we prefer the longest valid span boundary. Appendix A provides a one-page version of our guidelines with examples.

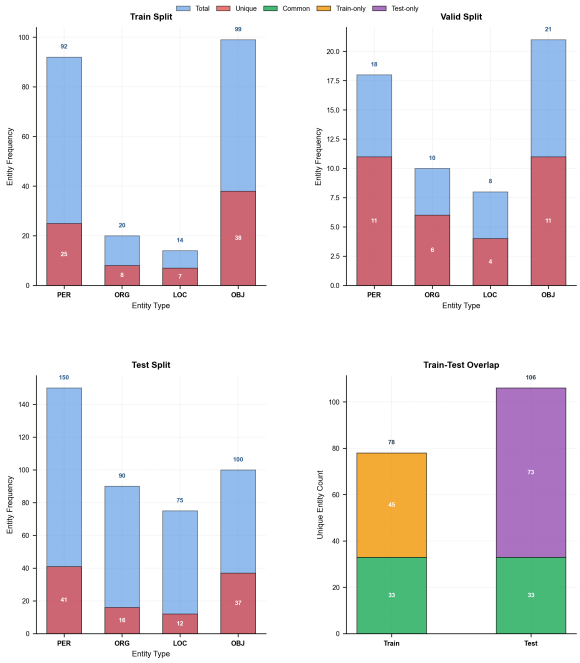

Figure 1: Dataset statistics for MythNER. The figure summarizes entity counts and label distributions across train/validation (P1) and test (P2), and reports the proportion of unique entities in the test set that overlap with training (31.1%).

4 Methods

4.1 Overview

We compare three families of approaches for PER/LOC/ORG/OBJ schema NER on MythNER: (i) a zero-shot spaCy baseline, (ii) a fine-tuned BERT token-classification baseline, and (iii) our proposed multi-agent LLM extraction system.

Our core hypothesis is that mythological narratives amplify the failure modes of off-the-shelf NER: long-range coreference, culturally specific aliases/titles, dense entity clusters, and heavy use of metaphorical or honorific expressions. The multi-agent design addresses these issues by (a) extracting with local context, (b) enforcing global consistency across chunks, and (c) verifying and selectively correcting uncertain spans.

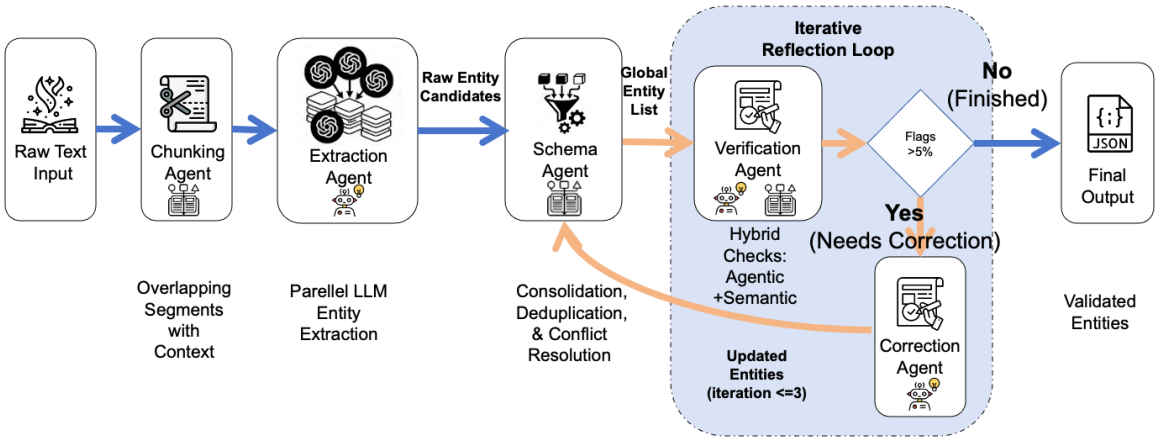

Figure 2: Multi-agent LLM extraction architecture. A text chunking agent creates overlapping segments with context; an extraction agent runs parallel LLM calls to propose entity spans; a consolidation agent deduplicates and resolves conflicts; a verification agent applies hybrid checks (agentic + semantic) and triggers an iterative reflection loop (up to 3 iterations) for targeted correction. The output is a validated entity set.

4.2 Proposed method: multi-agent LLM extraction

Workflow. The proposed system is a LangGraph-based pipeline[1] with five specialized roles: *chunking* (segmentation), *extraction* (LLM-based span proposal), *consolidation* (global merge across chunks), *verification* (QA over spans and labels), and *correction* (targeted re-extraction for flagged cases). Figure 2 depicts the following workflow:

> Input → Chunk → Extract → Consolidate → Verify → [Correct → Reverify] → Output.

Verification may trigger correction; the loop terminates early if verification passes, and is capped at 3 correction iterations.

Chunking with context windows. Given a full subtitle document, we split it into chunks of K characters and attach left/right context windows of C characters. This design makes extraction robust to boundary effects (e.g., a title introduced in one sentence and referenced in the next) while keeping each LLM call within a bounded context. Chunk boundaries are adjusted to nearby sentence-ending punctuation to reduce mid-phrase truncation.

Extraction prompting and confidence. The extraction stage queries an LLM with prompts specialized for (i) output format (document-level spans vs sentence-level spans), (ii) prompt type (a concise zero-shot prompt vs a richer prompt that includes guidelines and examples), and (iii) optional confidence reporting (continuous $[0, 1]$ or discrete 0–5 ranks).

Consolidation, verification, and targeted correction. Chunk-level proposals are merged into a single entity set for the document. A dedicated verifier then checks span validity (e.g., offsets are well-formed and aligned to the text), label plausibility (e.g., preventing obvious confusions such as locations mislabeled as persons), and cross-chunk consistency (e.g., repeated mentions of the same named entity). Only flagged items are sent to the correction stage, which performs targeted fixes rather than re-running the full extraction.

[1] LangGraph v1.0.5, MIT License. Homepage: https://docs.langchain.com/oss/python/langgraph/overview

Reproducibility knobs. The LLM used is Gemini-2.5-pro We record the chunk/context settings (K, C), prompt type, confidence mode, parallelism, and the LLM model identifier used in each run. Prompt templates are provided in Appendix B. These knobs are used in our controlled studies and ablations.

4.3 Baselines

4.3.1 Zero-shot spaCy

We use the pretrained zh_core_web_lg spaCy NER model without fine-tuning. Because spaCy's label set does not match our schema, we apply a deterministic mapping into {PER, LOC, ORG, OBJ}. Unmapped spaCy labels (e.g., DATE/TIME/QUANTITY) are ignored. Predictions are produced at the sentence level.

4.3.2 Fine-tuned BERT

We fine-tune bert-base-chinese for token classification with BIO tagging. The training label set is: {O, B/I-PER, B/I-LOC, B/I-ORG, B/I-OBJ}.

Training. HuggingFace Trainer is used for BERT model fine-tuning. We fix 3 random seeds (20261, 12345, 54321) and train each run for up to 10 epochs with early stopping (patience 3), learning rate 2×10^{-5}, weight decay 0.01, and batch size 4 with gradient accumulation 2 (effective batch size 8). The performance reported in Table 1 corresponds to the average of the best-performing checkpoint from each of the three runs.

Decoding to spans. At inference time, we decode token-level BIO predictions into entity spans using tokenizer offset mappings, producing sentence-local character spans in the coarse label set.

5 Experimental Setup

5.1 Preprocessing

Common format. Gold annotations are stored as JSONL with sentence-level texts, entity offsets, and entity labels (PER/LOC/ORG/OBJ).

Label mapping for spaCy. We map spaCy default labels to coarse labels via:

PERSON→PER;
{FAC,GPE,LOC}→LOC;
{ORG,NORP}→ORG;
{PRODUCT,WORK_OF_ART,EVENT}→OBJ.

Unmapped spaCy labels (e.g., DATE/TIME/QUANTITY) are ignored.

Data splits and standardized format. We split *Ne Zha* P1 subtitles for train/validation (80/20) and P2 as a held-out test set. To create baseline training splits, we convert JSONL annotations into the spaCy format to standardize model training and evaluation. We use a leakage-aware strategy: sentences are grouped into connected components based on co-occurring entity strings, and entire components are assigned to train vs dev. This reduces leakage from repeated names/aliases across splits.

BERT data conversion. We further convert the spaCy format into BERT BIO tags using tokenizer offset mappings. Special tokens [CLS] and [SEP] are assigned loss label -100. In training/evaluation we truncate to a maximum sequence length of 64 tokens.

5.2 Evaluation

We report exact span-match micro-averaged precision/recall/F1 over entity tuples (start, end, label). A predicted entity is counted as correct if and only if its span boundaries and label exactly match a gold entity in the same sentence. We micro-average by summing true positives, false positives, and false negatives across all sentences.

All methods are evaluated in a unified representation as sentence-local character spans; if a method produces document-level offsets internally, they are mapped back to sentence-local offsets before scoring.

6 Results

6.1 Main results

Table 1 compares a general-purpose Chinese spaCy model (zero-shot), a supervised BERT token-classification baseline, and our multi-agent extraction pipeline.

The zero-shot spaCy baseline substantially underperforms (F1=0.185), reflecting the domain shift in mythological narratives. Fine-tuning a strong encoder improves performance (BERT-base-chinese: F1=0.655), while the multi-agent pipeline achieves the best overall

Model	P	R	F1
Zero-shot spaCy	0.271	0.141	0.185
Fine-tuned BERT	0.613	**0.704**	0.655
Multi-agent LLM	**0.770**	0.691	**0.728**

Table 1: Main NER results on *Ne Zha*. We report exact span match micro-P/R/F1 on the held-out *Ne Zha* P2 test set.

Chunk	Context	P	R	F1
200	250	0.779	0.667	0.718
250	200	0.770	**0.691**	**0.728**
1000	500	0.768	0.496	0.603
2000	1000	0.773	0.407	0.533
4000	1000	**0.788**	0.310	0.445

Table 2: Chunk/context tuning for the multi-agent pipeline (few-shot+guide with extracted-entity examples). We report exact span match micro-P/R/F1 on the held-out *Ne Zha* P2 test set.

Setting	Variant	P	R	F1
Prompting				
	Zero-shot (continuous)	0.459	**0.822**	0.589
	Few-shot + guide (continuous)	**0.770**	0.691	**0.728**
Confidence Scoring				
	Few-shot + guide (continuous)	**0.770**	0.691	**0.728**
	Few-shot + guide (discrete)	0.758	**0.697**	0.726
Example Richness				
	Few-shot + guide (discrete)	**0.758**	0.697	**0.726**
	Fully annotated (discrete)	0.562	**0.757**	0.645

Table 3: Ablation slices for the multi-agent pipeline at the best chunk/context setting. We separate prompting style, confidence scoring, and example richness for readability.

result (F1=0.728), improving over BERT by +0.073 F1 and over the zero-shot baseline by +0.543 F1.

6.2 Configuration study

We next analyze which design choices drive performance (Table 2; Table 3).

Chunk/context tuning. With the prompting strategy held fixed (few-shot + guide using extracted-entity examples), moderate chunk and context sizes yield the strongest results; the best configuration is a chunk size of 250 with a context window of 200.

Prompting, confidence, and example richness. Using only label definitions (agent zero-shot) yields high recall but much lower F1 than few-shot+guide prompting. Confidence scoring (continuous vs discrete) has a small effect at the best chunk/context. For example richness, we observe that fully annotated examples underperform simple few-shot examples in our current prompting setup; we treat this as a prompt-formatting effect rather than a definitive statement about annotation richness.

7 Analysis and Discussion

7.1 The challenge of myth-domain NER

Myth-domain NER is challenging because our label space is defined around *referable instances* rather than surface words. Following our annotation criteria, a span should be annotated only if it can answer the core questions *"who / what / where"*. Consequently, many expressions that look "important" in myth narratives are deliberately excluded, including events or phenomena such as *"天劫"* (*heavenly tribulation*) and *"封神大战"* (*Fengshen War*).

The most ambiguous label is OBJ. In our setting, OBJ is reserved for *named* artifacts or fixed technique titles (e.g., a spell name invoked as a unit), while generic concepts such as *"命运/因果/力量"* (*fate/karma/power*) are not annotated unless they behave as a proper name in context. This "named-object" constraint is crucial for dataset consistency, but it also makes the boundary between "object" and "non-entity" particularly sharp and error-prone.

Subtitles further amplify ambiguity through pervasive titles and references. Generic titles like *"师尊/师弟"* (*master/junior disciple*) are excluded, while unique, referential variants such as *"李大人"* (*Lord Li*), *"吒儿"* (*Zha'er, a nickname*), and *"三公子"* (*third young master*) may be annotated as PER when they uniquely identify a character.

Finally, we adopt longer span preference : when a full mention is available, we annotate the complete name (e.g., *"乾元山金光洞太乙真人"* / *Taiyi Zhenren of Qianyuan Mountain*

Jinguang Cave). While this reduces fragmentation in the gold labels, it increases the probability of exact-span mismatches for models.

This design choice interacts directly with our evaluation protocol. Because we use exact span match, boundary near-misses are penalized maximally: a prediction that captures the right referent but trims or extends a title-heavy mention is still counted as an error. In practice, this makes span normalization a central challenge in MythNER, and it partly explains why boundary-related errors dominate across both conventional taggers and LLM-based systems.

7.2 Traditional baseline errors

We observe several recurring failure modes in traditional baselines. First, zero-shot models frequently over-predict "entity-like" common words, producing false positives from adjectives, verbs, and generic nouns that do not satisfy the *referable-instance* criterion. Second, both zero-shot and supervised baselines are sensitive to long, title-heavy mentions, often fragmenting a gold span into smaller pieces. Under exact-span evaluation, these near-misses are counted as errors even if a partial substring is correct.

Type confusion is also prominent in mythology-specific contexts where the same surface form can plausibly be a person-like being, a collective, or a named object. For example, role-title composites and faction-like phrases can trigger drift between PER/ORG/LOC. Finally, OBJ remains a residual bottleneck: models may miss named objects/techniques entirely (recall errors) or incorrectly promote abstract substances or generic items to OBJ (precision errors), reflecting the intrinsic ambiguity of "named artifact vs. generic concept" in myth discourse.

7.3 LLM-agent errors: what changes and what remains

The LLM-agent pipeline changes the error profile rather than eliminating errors altogether. In an agent *zero-shot* configuration (high-recall, lightly constrained), the dominant issue is over-extraction: the agent tends to label abstract substances and generic nouns as OBJ, e.g., "天地灵气" (*heaven-and-earth aura*), "仙气" (*immortal aura*), and "魔气" (*demonic aura*), which are not annotated in the gold standard. It also exhibits precision collapse for PER by eagerly labeling context-dependent addresses (e.g., "猪兄" / *Brother Pig*, "仙长" / *immortal elder*) that our dataset often treats as non-entities.

Meanwhile, exact-span sensitivity remains: long mentions are frequently split into overlapping subspans (a typical false-positive/false-negative pair under span-level matching). Even in the extracted, tuned chunk/context settings, we still observe boundary drift where the agent extends a named object with nearby descriptive material (e.g., predicting "宝莲仙气" / *lotus aura* instead of the gold "宝莲" / *lotus*).

Compared to agent zero-shot, tuned extraction configurations drastically reduce false positives but increase false negatives, illustrating a clear precision–recall tradeoff. In practice, many of the remaining misses concentrate on short, frequently mentioned myth objects/techniques such as "魔丸" (*Demon Pill*) and "天雷" (*heavenly thunder*), which require stable span decisions across fragmented subtitle context.

The configuration study in our results further supports a simple principle: context helps until it hurts. Too little context deprives the model of cross-mention cues for disambiguation, while too much context (or overly large chunks) increases boundary drift and encourages "entity-like" generic nouns, especially for OBJ. This reinforces the role of constraint design and iterative verification in agentic extraction.

7.4 Practical takeaway

Across baselines, the dominant challenges are (i) **boundary alignment** under exact-span scoring, (ii) **type ambiguity** in mythology-specific language, and (iii) the particularly sharp **scope constraint** of OBJ (named artifact/technique vs. generic concept). The LLM-agent approach reduces some forms of type drift via stronger contextual reasoning, but introduces *over-extraction* when constraints are weak and remains highly sensitive to span normalization, motivating careful confidence scoring and post-processing for deployment.

Overall, MythNER is less a test of mem-

orizing surface patterns than a test of deciding entityhood and stabilizing boundaries under culturally grounded naming conventions. The dataset therefore complements existing benchmarks by emphasizing (i) conservative referable-instance criteria, (ii) long-span boundary decisions, and (iii) ontology-like constraints implicit in OBJ.

8 Limitations

MythNER targets a specific long-tail setting: Chinese mythological narratives in subtitle form. As a result, models trained and evaluated on MythNER may not transfer directly to other genres (e.g., classical prose, novels) or other media with different discourse conventions.

Our annotation design also imposes deliberate scope constraints. We adopt a flat, non-overlapping schema with four coarse types (PER/LOC/ORG/OBJ), and we use conservative entityhood criteria that exclude many events and phenomena. In particular, OBJ is restricted to *named* artifacts or fixed technique titles; this improves consistency but creates a sharp boundary between named objects and generic concepts.

Evaluation uses exact span match, which is appropriate when downstream use requires boundary-faithful extraction, but it penalizes near-miss boundary predictions maximally. This interacts with our longer-span preference and makes span normalization a primary source of error.

Finally, our LLM-agent pipeline introduces practical constraints: performance is sensitive to prompting and chunk/context choices, and deployment may incur higher cost and latency than conventional taggers.

References

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuanjing Huang. 2019. A Lexicon-Based Graph Neural Network for Chinese NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1040–1050, Hong Kong, China. Association for Computational Linguistics.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on Named Entity Recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Bin Li, Bolin Chang, Ruilin Liu, Xue Zhao, Si Shen, Lihong Liu, Yan Zhu, Zhixing Xu, Weiguang Qu, and Dongbo Wang. 2025. Overview of EvaHan2025: The First International Evaluation on Ancient Chinese Named Entity Recognition. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 156–164, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Kalyani Pakhale. 2023. Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges. *Preprint*, arXiv:2309.14084.

Siyao Peng, Zihang Sun, Sebastian Loftus, and Barbara Plank. 2024. Different Tastes of Entities: Investigating Human Label Variation in Named Entity Annotations. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 73–81, Malta. Association for Computational Linguistics.

Aravinth Sivaganeshan and Nisansa De Silva. 2023. Fine Tuning Named Entity Extraction Models for the Fantasy Domain. In *2023 Moratuwa Engineering Research Conference (MERCon)*, pages 346–351.

Xinli Tao, Xin Dong, and Xuezhong Zhou. 2025. OEMA: Ontology-Enhanced Multi-Agent Collaboration Framework for Zero-Shot Clinical Named Entity Recognition. *Preprint*, arXiv:2511.15211.

Simone Tedeschi and Roberto Navigli. 2022. Multi-NERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025a. 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.

Zihan Wang, Ziqi Zhao, Yougang Lyu, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2025b. A Cooperative Multi-Agent Framework for Zero-Shot Named Entity Recognition. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 4183–4195, New York, NY, USA. Association for Computing Machinery.

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.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arık. 2024. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237.

Yufan Zhuang, Xiaodong Yu, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Jingbo Shang, Zicheng Liu, and Emad Barsoum. 2025. Self-Taught Agentic Long Context Understanding. *Preprint*, arXiv:2502.15920.

A Annotation Guide

We annotate *flat*, non-overlapping named entities in Chinese mythological subtitles with four coarse types: PER, LOC, ORG, OBJ. The annotation target is not "words", but *referable instances in context*.

Entityhood test (fast rule). Annotate a span only if, in its local context, it can reasonably answer *who / what (thing) / where*. If an expression only answers "what event/state/phenomenon is happening", do *not* annotate it.

Labels. PER: a specific character or person-like being (humans, immortals, demons, named creatures). LOC: a named place/location. ORG: a named faction/organization acting as a collective. OBJ: a *named* artifact/item, or a fixed technique title invoked as a unit.

Span rules. (1) **No overlaps / no nesting** (flat NER). (2) **Prefer the longest valid boundary** when a full canonical mention is available. (3) Exclude pure pronouns and generic role nouns.

Titles, aliases, and references. Generic, context-dependent addresses (e.g., 师尊/师兄/大人) are not annotated. However, if a variant uniquely identifies a character (contains a surname, name fragment, rank/epithet that is used as a unique handle), annotate it as PER.

OBJ is conservative by design. Annotate OBJ only when the mention behaves like a proper name or a stable technique title. Do not annotate generic substances/abstract concepts (e.g., 神力/力量/因果/命运) unless they are explicitly treated as a named unit in context.

Exclude events and phenomena. Named or salient events/rituals/astronomical phenomena are not entities (e.g., 天劫, 封神大战, 生辰宴).

LOC vs. ORG (surface-syntax heuristic). Some names can denote both a place and an institution. We default to LOC for consistency.

Examples.

> **PER (unique name).** 我乃 [太乙真人] PER。 [李大人] PER 来了。 但是 [师尊] *not annotated.*
>
> **LOC (named place).** 我们去 [陈塘关] LOC。 他在 [昆仑山] LOC 修行。
>
> **ORG (faction acting collectively).** [天庭] ORG 下令捉拿。 [妖族] ORG 来犯。
>
> **OBJ (named artifact / technique title).** 哪吒祭出 [乾坤圈] OBJ。 他使出 [龙族秘术] OBJ。 但 [神力/力量] *not annotated.*
>
> **Long-span preference.** [乾元山金光洞太乙真人] PER (not 乾元山 + 金光洞 + 太乙真人).
>
> **Exclude events/phenomena.** [天劫] *not annotated*。 [封神大战] *not annotated*。

B LLM Prompt Templates

This appendix documents the prompt templates used by the multi-agent pipeline. We report (i) the prompts by agent role and (ii) the prompt variants used in our ablation settings (Table 3). Templates contain placeholders such as {chunk_text} and {start_char}.

B.1 Agent prompts (by role)

Chunking agent (no LLM call). Chunking is programmatic: a document is split into chunks of length K and each chunk is provided with left/right context windows of length C. These parameters change the *inputs* to the extraction prompt (the chunk text and its context), but do not change the prompt text itself.

Extraction agent: zero-shot, label definitions only. Original (ZH).

> 你是一个命名实体识别 (NER) 专家。
>
> 任务：从中文神话文本中提取命名实体，遵循以下 schema：
>
> 【实体类型定义】- PER: 人物，包括人类、神仙、妖魔、龙族等能说话、行动的角色。- LOC: 地点，具体的地理位置或场所名称。- ORG: 组织，有组织性的群体、种族、派系、领域。- OBJ: 物品，法器、武器、法术名、魔法物品、坐骑等名。
>
> 输出格式：必须返回有效的 JSON：{ "entities": [{ "text$_{span}$" : " ","$start_{char}$" : 0,"end_{char}" : 2,"*type*" : "*PER*","*confidence*" : 1.0}]}

重要：start_char 和 end_char 是相对于完整原文档的绝对位置！

—

请从以下【文本片段】中提取所有命名实体。

【文本片段】（第 {start_char} 到 {end_char} 字符）：{chunk_text}

【上文】（仅供参考）：{context_before}

【下文】（仅供参考）：{context_after}

提取要点：1. 只从【文本片段】提取，位置必须是文档绝对位置。2. 只要输出 entities 列表。

English rendering.

> You are an expert in Named Entity Recognition (NER).
>
> Task: Extract named entities from Chinese mythology text, using this schema: - PER: characters/person-like beings. - LOC: named locations. - ORG: named organizations/factions. - OBJ: named artifacts/weapons/spell or technique titles/mounts.
>
> Output must be valid JSON: { "entities": [{ "text$_{span}$" : "*ENTITY*","$start_{char}$" : 0,"end_{char}" : 2,"*type*" : "*PER*","*confidence*" : 1.0}]}

IMPORTANT: start_char/end_char are absolute offsets in the full document.

—

Extract all named entities from the following [Text Fragment] only. [Text Fragment] (characters {start_char} to {end_char}): {chunk_text}

[Previous Context] (reference only): {context_before}

[Following Context] (reference only): {context_after}

Extraction agent: few-shot + guide; continuous confidence. This variant adds explicit boundary/type rules and in-prompt examples, and requests a continuous confidence score $\in [0, 1]$.

Original (ZH).

> 你是一个中国神话文本的命名实体识别（NER）专家。
>
> 任务：从中文神话文本中提取命名实体，遵循 Stage 1 schema：
>
> 【实体类型定义】- PER（人物）：能说话、行动的角色，包括人类、神仙、妖魔、龙族 包括：太乙真人、哪吒、敖丙、李靖、申公豹、龙王、海夜叉、咤儿、太乙、天尊 包括带地点的完整称谓："乾元山金光洞太乙真人"整体是 PER 不包括：通用称呼（仙长、师父）、代词、"X 夫妇"中只标注人名
>
> - LOC（地点）：具体地理位置或场所名称 包括：陈塘关、龙宫、昆仑山、东海、乾元山、金光洞、海底炼狱 不包括：天庭、仙界（组织/领域）、方位词（南面、北面）

- ORG（组织）：有组织性的群体、种族、派系、领域 包括：龙族、天庭、十二金仙、妖族、阐教、仙界、伏魔帮、人族

- OBJ（物品）：法器、武器、法术名、魔法物品、坐骑 包括：混元珠、灵珠、魔丸、乾坤圈、火尖枪、混天绫、天劫咒、结界、虚空之门、风火轮 不包括：咒语念词（如"急急如律令"）

【关键提取规则】1. 边界规则：提取核心实体名，不要过度扩展 2. 只提取文本片段中的实体 3. 同一实体的不同形式都要提取 4. 不标注：通用称呼、方位词、代词、咒语念词

输出格式：必须返回有效的 JSON：{ "entities": [{ "text$_{span}$" : " ","$start_{char}$" : 45,"end_{char}" : 47,"$type$" : "PER","$confidence$" : 0.95}]}

重要：start_char 和 end_char 是文档绝对位置！

—

【文本片段】（第 {start_char} 到 {end_char} 字符）：{chunk_text}

【上文】（参考）：{context_before}

【下文】（参考）：{context_after}

提取要点：1. 只从文本片段提取 2. 区分：天庭/仙界 =ORG；陈塘关/龙宫 =LOC；结界/虚空之门 =OBJ 3. 提取核心实体名

English rendering.

You are an expert NER annotator for Chinese mythology text.

Task: Extract named entities under a Stage-1 schema (PER/LOC/ORG/OBJ).

Boundary rule: extract the core entity name; do not over-extend spans.

Output must be valid JSON with absolute offsets: { "entities": [{"text$_{span}$" : "...","$start_{char}$" : 0,"end_{char}" : 2,"$type$" : "PER","$confidence$" : 0.95}]}

Extract entities from the Text Fragment (characters {start_char} to {end_char}): {chunk_text}

Extraction agent: discrete confidence variant (0–5 ranks). Original (ZH).

置信度等级（0–5）：- 5: 绝对确定 - 4: 高置信度 - 3: 中等置信度 - 2: 低置信度 - 1: 极低置信度 - 0: 不确定

输出格式：{ "entities": [{ "text$_{span}$" : " ","$start_{char}$" : 45,"end_{char}" : 47,"$type$" : "PER","$confidence_{rank}$" : 5,"$confidence_{reason}$" : " "}]}

追加要求：每个实体提供 confidence_rank 与 confidence_reason。

English rendering.

Discrete confidence (0–5). Output JSON must include: { "entities": [{ "text$_{span}$" : "...","$start_{char}$" : 0,"end_{char}" : 2,"$type$" : "PER","$confidence_{rank}$" : 5,"$confidence_{reason}$" : "$shortjustification$"}]}

Extraction agent: fully annotated examples. Original (ZH).

输出格式：JSON 数组，每句一个对象：[{ "id": 1, "text": " 这就是我万人敬仰的太乙真人", "label": [[9, 13, "PER"]], "Comments": [] }]

重要：label 偏移量相对于句子 text！

English rendering.

Output format: JSON array, one object per sentence. Offsets are relative to the sentence text.

Verification agent. Original (ZH).

你是一个 NER 标注质量检查专家。

任务：验证实体提取是否符合 Stage 1 规则：1. 非重叠 2. 最大跨度 3. 类型正确 4. OBJ 必须是专有名称 5. 边界准确

输出格式：{ "summary": {"total$_{e}ntities$" : 0,"$approved_{c}ount$" : 0,"$flagged_{c}ount$" : 0,"$issues$" : []},"$detailed_{f}indings$" : [{"$entity_{i}d$" : "...","$issue_{t}ype$" : "$boundary_{e}rror$","$description$" : "...","$suggested_{f}ix$" : "..."}]}

原文（前 5000 字符）：{raw_text} 提取的实体：{entities_json}

English rendering.

You are a quality-control agent for NER. Return JSON with summary and detailed findings.

Correction agent. Original (ZH).

你是一个 NER 标注修正专家。

任务：根据验证报告修正实体。

输出格式：{ "corrections": [...], "rebuttals": [...] }

被标记的实体：{flagged_entities_json} 验证报告：{verification_report_json} 扩展上下文：{expanded_contexts} 原文（前 10000 字符）：{raw_text}

English rendering.

You are a correction agent. Apply fixes or rebuttals and return JSON.

B.2 Prompt variants used in ablations

Prompt type. *Zero-shot* uses concise label definitions only. *Few-shot + guide* adds explicit rules and examples.

Confidence scoring. *Continuous*: real-valued confidence in $[0, 1]$. *Discrete*: integer confidence_rank plus confidence_reason.

Example richness. *Extracted-entity examples*: entity-level prompts. *Fully annotated examples*: sentence-level labeled format.

Chunk/context. Chunk/context settings (K, C) determine the injected chunk_text and context_before/after.