Beyond Boundaries: Learning a Universal Entity Taxonomy across Datasets and Languages for Open Named Entity Recognition

Yuming Yang[1], Wantong Zhao[1], Caishuang Huang[1], Junjie Ye[1], Xiao Wang[1], Huiyuan Zheng[1], Yang Nan[1], Yuran Wang[2], Xueying Xu[2], Kaixin Huang[2], Yunke Zhang[2], Tao Gui[3,4*], Qi Zhang[1,5,6*], Xuanjing Huang[1,6*]

[1] School of Computer Science, Fudan University [2] Honor Device Co., Ltd
[3] Institute of Modern Languages and Linguistics, Fudan University [4] Pengcheng Laboratory
[5] Research Institute of Intelligent Complex Systems, Fudan University
[6] Shanghai Key Laboratory of Intelligent Information Processing

yumingyang23@m.fudan.edu.cn {qz,tgui,xjhuang}@fudan.edu.cn

Abstract

Open Named Entity Recognition (NER), which involves identifying arbitrary types of entities from arbitrary domains, remains challenging for Large Language Models (LLMs). Recent studies suggest that fine-tuning LLMs on extensive NER data can boost their performance. However, training directly on existing datasets neglects their inconsistent entity definitions and redundant data, limiting LLMs to dataset-specific learning and hindering out-of-domain adaptation. To address this, we present B^{2}NERD, a compact dataset designed to guide LLMs' generalization in Open NER under a universal entity taxonomy. B^{2}NERD is refined from 54 existing English and Chinese datasets using a two-step process. First, we detect inconsistent entity definitions across datasets and clarify them by distinguishable label names to construct a universal taxonomy of 400+ entity types. Second, we address redundancy using a data pruning strategy that selects fewer samples with greater category and semantic diversity. Comprehensive evaluation shows that B^{2}NERD significantly enhances LLMs' Open NER capabilities. Our B^{2}NER models, trained on B^{2}NERD, outperform GPT-4 by 6.8-12.0 F1 points and surpass previous methods in 3 out-of-domain benchmarks across 15 datasets and 6 languages. The data, models, and code are publicly available at https://github.com/UmeanNever/B2NER.

1 Introduction

Open Named Entity Recognition (NER), which targets both in-domain and out-of-domain identification of common and unseen entities, is crucial for broader NER applications in real-world scenarios, such as the low-resource fields of law and biomedicine (Etzioni et al., 2008; Leitner et al., 2019; Perera et al., 2020). As shown in Figure 1, despite advancements in Large Language Models

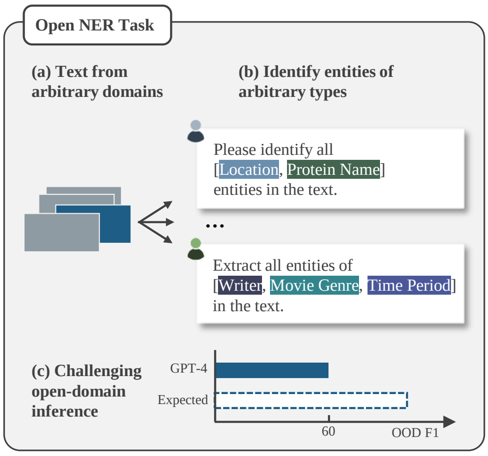


Figure 1: The Open NER task aims to extract arbitrary entities (common and unseen) from arbitrary domains (in-domain and out-of-domain). Current LLMs, like GPT-4, still fall short on this task.

(LLMs) raising expectations for solving Open NER, current LLMs still struggle with intricate entity taxonomies in open domains and show limited NER capabilities (Katz et al., 2023; Gao et al., 2023; Wei et al., 2023; Li et al., 2023; Ye et al., 2023). Recent studies (Wang et al., 2023b; Sainz et al., 2024; Xiao et al., 2023; Gui et al., 2024) address this by fine-tuning LLMs on numerous existing NER datasets, helping them learn detailed entity definitions and achieve better overall performance.

However, directly using existing datasets to train Open NER models is hindered by two flaws that limit the models' out-of-domain generalization: (1) **Inconsistent and vague entity definitions across datasets.** Different datasets often have conflicting entity definitions. For instance, some datasets distinguish between locations like "Times Square" and geopolitical entities like "Paris", while others annotate both as LOC. Aligning LLMs with these inconsistencies leads to dataset-specific patterns and confusion over common entities during inference (Figure 2). To avoid conflicts, Zhou et al., 2024 sug-

*Corresponding Authors.

Proceedings of the 31st International Conference on Computational Linguistics, pages 10902–10923
January 19–24, 2025.

gests adding dataset names in training prompts, but this cannot improve out-of-domain inference with unknown datasets. Sainz et al., 2024 introduces detailed annotation guidelines for each entity type, but such guidelines are hard to obtain and challenging for LLMs to understand. (2) **Redundant data in combined datasets.** Most datasets heavily annotate common entities, with fewer samples for long-tail entities. Thus, the combined dataset contains redundant samples with similar annotations and semantics. This lack of diversity may cause LLMs to overfit and hinder universal generalization (Zhou et al., 2023; Liu et al., 2024). To circumvent above issues, some studies (Zhou et al., 2024; Li et al., 2024; Ding et al., 2024) explore using synthetic NER data annotated by ChatGPT, but synthetic data struggles to meet real-world NER requirements. The valuable human annotations in existing datasets remain underutilized.

In this work, we propose enhancing LLMs for Open NER by directly addressing issues in existing training datasets and normalizing them into a compact collection via a two-step approach. First, we systematically standardize entity definitions across all collected datasets. Inconsistent entity definitions are automatically detected via model-based validation and rule-based screening. We then clarify these ambiguous entity types by assigning distinguishable label names for each unique type following detailed principles. This step forms our universal entity taxonomy, which guides the categorization of both common and unseen entities. Second, we avoid redundancy by employing a data pruning strategy that considers both category and semantic diversity. Our strategy samples equally from each entity type and selects samples with lower textual similarity within each type to enhance semantic diversity. By applying above approach on our bilingual (English and Chinese) NER collection of 54 datasets, we derive B^{2}NERD, a **B**eyond-**B**oundary **NER D**ataset with a universal taxonomy of 400+ entity types across 16 major domains.

By fine-tuning on B^{2}NERD, we develop B^{2}NER models — LLMs with extensive Open NER capabilities that generalize across datasets and languages. Experimental results on 3 out-of-domain NER benchmarks across 15 datasets show that our model outperforms both GPT-4 and previous methods by 3.0% in English, 6.8% in Chinese, and 6.7% in a multilingual setting. Further analysis offers deeper insights into our approach's effectiveness.

Our main contributions are three-fold:

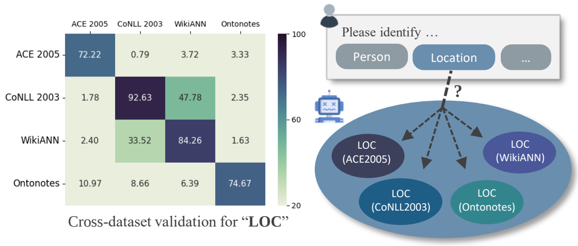


Figure 2: Sample results of BERT-based cross-dataset entity validation for LOC entity. Light colors indicate conflict entity definitions. Training LLM on these inconsistent datasets leads to confusions during inference.

- We present B^{2}NERD, a cohesive and compact dataset that advances LLM capabilities for Open NER, along with its full version, B^{2}NERD$_{\text{all}}$, the largest bilingual NER data collection to date.
- We introduce a two-step approach to address the inconsistencies and redundancy among existing NER datasets, creating a universal entity taxonomy that transcends dataset boundaries.
- Experiments show that our B^{2}NER models outperform GPT-4 and previous methods in comprehensive out-of-domain evaluations across various datasets and languages.

2 Preliminaries

We first discuss our data collection process and the limitations of using collected datasets.

2.1 Data Collection

To meet the diverse needs of Open NER, we gather the largest collection of existing datasets. For English NER, we use the collection from Wang et al., 2023b. For Chinese NER, we invest extensive effort in data collection due to the limited datasets in prior work. More details are in Appendix A.1. Finally, we derive a bilingual collection of 54 datasets from 16 major domains, as shown in Figure 4.

2.2 Inconsistencies among Collected Datasets

The collected datasets have differing entity definitions. To quantify their conflicts, we conduct cross-dataset entity validation experiments. Figure 2 shows a sample experiment among 4 datasets, all having LOC entities. We iteratively train a BERT-based model on one dataset and evaluate its performance on recognizing LOC entities in the other three. Low F1 scores indicate inconsistent entity definitions between datasets. The results re-

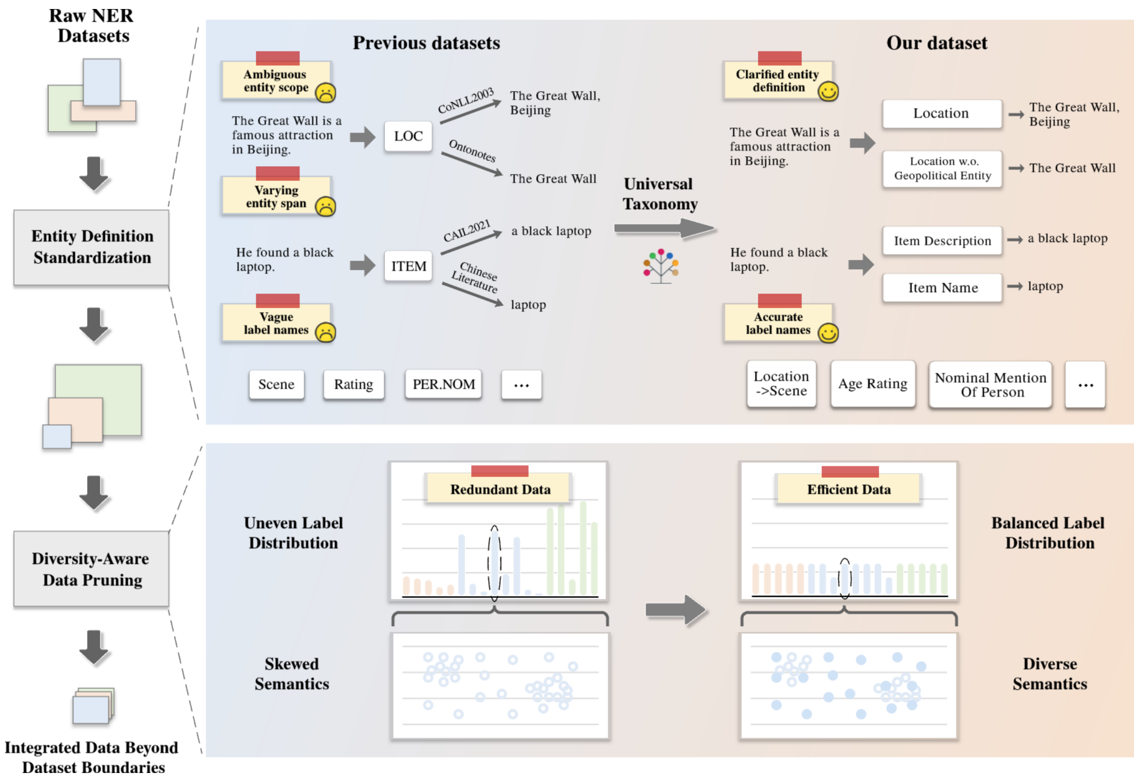


Figure 3: Framework of B^{2}NERD data construction: raw NER datasets are reshaped into a cohesive dataset via entity definition standardization and diversity-aware data pruning. Final data is then used to train our Open NER model.

veal significant conflicts among collected datasets, which confuse LLMs during training and inference. More explanations are provided in Appendix A.2.

3 Approach

We propose a two-step approach to address existing dataset inconsistencies and redundancy. This section details the construction of the B^{2}NERD dataset (Sections 3.1-3.2) and the training of B^{2}NER models (Section 3.3). Figure 3 outlines our framework.

3.1 Entity Definition Standardization with Universal Taxonomy

As shown in Figures 2 and 3, the same entity label often has different meanings across datasets, and many labels are unclear outside their original context. To address these ambiguities and avoid dataset-specific learning, we systematically standardize entity definitions in existing datasets by detecting conflicts, clarifying ambiguous entities for a universal taxonomy, and renaming entity labels. New entities can also be easily accommodated within this taxonomy following our practice.

Automatic dataset conflict detection. First, we detect conflicts among datasets at scale by identifying inconsistent annotations for entities with similar label names using automatic methods: **Model-based cross validation**: We extend the method in Section 2.2 to all dataset pairs with similar entity types, identifying potential conflict entity definitions from low F1 results. **Rule-based screening**: To further understand these conflicts, we screen for cases when same entity mention receives different annotations across datasets. Significant inconsistent cases are classified and listed for future processing. See Appendix A.2 for more details.

Resolving Conflicts and Constructing Universal Taxonomy. For entities with similar label names but different definitions, we invite experts to scrutinize their differences and split them into unique entity types, following NER guidelines like ACE[1]. As partially shown in Figure 3, we address major issues like: (1) **Different entity scope**: The same label name might encompass a different range of entities, as shown in the LOC example in Figure 3. (2) **Different entity span**: Different datasets may identify different spans for the same entity label, as shown in the ITEM example in Figure 3. (3) **Different mention type**: There are various ways to refer an entity. For PER entities, most datasets recog-

[1] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

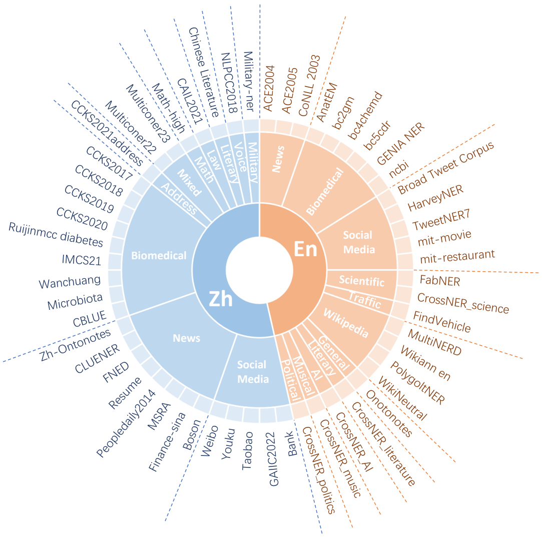


Figure 4: Overview of collected datasets in B^{2}NERD. We collected 54 datasets across 16 major domains.

nize explicit names like "Jack Smith", while others, like ACE 2005, include nominal or pronoun mentions like "a hunter" and "he" as PER. The latter will be split as a new entity type GENERAL MENTIONS OF PERSON. (4) **Overlapped entity granularity**: When a dataset contains both coarse and fine-grained entities, like PERSON and WRITER, the model may only label the coarse type (Sainz et al., 2024). We believe the former type actually refers to "other person" in such datasets and should be distinguished as PERSON->OTHERS. By applying this clarification and separation process to each entity type, we create a universal entity taxonomy with consistent definitions across all datasets.

Reassigning Label Names. Label names are crucial in generative NER as they appear in both prompts and answers, depicting current task. Thus, we reassign natural language labels to entity types in the universal taxonomy to more accurately represent their definitions for LLM understanding. Our naming system adheres to the following principles: **Readable**: Labels should be clear words or phrases, avoiding acronyms. **Unambiguous**: Each label should distinctly differentiate between similar entity types, like ITEM NAME v.s. ITEM DESCRIPTION. **Hierarchical**: Entity sub-types are named after their parent types (e.g., LOCATION->SCENE), aiding in polysemous resolution and clear granularity levels. **Flexible**: To accommodate diverse NER tasks and new entities types, the system allows adaptable naming, such as using "or" in labels. For example, a type encompassing both person and group entities is labeled PERSON OR GROUP.

The clarifying and renaming process is performed by human experts to ensure higher data quality. We publicly recruit college students with sufficient NER annotation experience and implement procedures such as preliminary training and attention checks to maintain overall work quality and consistency (see Appendix A.3 for details). The final taxonomy includes 400+ diverse entity types, as shown in Table 1. The standardized datasets with consistent entities comprise our B^{2}NERD$_{\text{all}}$ collection, containing the full data.

3.2 Diversity-aware Data Pruning

Despite addressing inconsistencies, the merged dataset retains imbalanced data distribution from raw datasets. For instance, LOCATION entities in news are heavily annotated, while long-tail entities like CITY are sparse. To avoid model over-fitting to redundant data, we propose fine-tuning LLMs on a curated, diverse subset of B^{2}NERD$_{\text{all}}$ to learn more transferable patterns.

As depicted in the lower part of Figure 3, we address diversity in NER datasets by ensuring **balanced data distribution** across entity types and **diverse text semantics** within each type's samples.

To maximize diversity in a limited-size dataset, our strategy selects k semantically diverse samples per unique entity type. We start by initializing sample pools for all entity types $S_1, \cdots, S_M = \emptyset$, each holding up to k samples. Then for each random sample x with annotations of entity type $(t_1, t_2, ...t_M)$, we check the status of related pools $(S_{t_1}, S_{t_2}, ..., S_{t_M})$. For non-full pools, we assess the maximum semantic similarity between x and corresponding sample pool S_{t_i}. We decide whether to add x to this pool based on the probability:

$$p(S_{t_i} \leftarrow S_{t_i} \cup \{x\}) = 1 - \max_{y \in S_{t_i}} \text{sim}(x, y)$$
$$= 1 - \max_{y \in S_{t_i}} \text{cosine}(x, y) + b$$

The offset b controls the penalty for semantic similarity. We utilize AnglE (Li and Li, 2023) to generate text embeddings for samples and calculate semantic distance using cosine similarity between their embeddings. This approach prioritizes semantically diverse samples. Notably, if a sample is added to a pool, all entity mentions within it are retained for optimal data efficiency. The sampling process continues until all pools are full or all samples have been traversed. The final selected dataset is formed by combining unique samples from all

Person	Location	Organization	Object	Creative Work
Person Person->Artist Person->Scientist Person->Victim General Mention of Person Nationality ...	Location Location->Facility Location->Starting point Geo-Political entity Location(w/o Country) City ...	Organization Organization->Sport team Organization->Corporation Organization->Political Group Organization->Band Organization->Public Company ...	Astronomical Object Product Name Product Name->Music Instrument Product Name->SUV Crime Tool Item Description ...	Creative Work Creative Work->Album Creative Work->Literature Creative Work->Magazine Creative Work->Software Music Theme ...
Measure	**Time**	**Education**	**Biomedical**	**Other**
Cardinal Number Dose Measurement Quantity Monetary Amount(with unit) Property Value Percentage(with %) ...	Date or Period Sub-day Time Expression Duration Generation Operating Hours Frequency ...	Academic Major Research Field Education Background Mathematical concept Mathematical principle or method Academic Conference ...	Anatomy Drug or Vaccine Chemical Microorganism Disease Name Symptom ...	Event Name Mobile Phone Model Legal Document Financial Term Type Of Emotion Color ...

Table 1: The universal entity taxonomy for B^{2}NERD includes 400+ entity types across 10+ main categories. New entities can be easily accommodated within this taxonomy. The full table is available in Appendix D.

Split	Lang.	Datasets	Types	Num	Raw Num
Train	En	19	119	25,403	838,648
	Zh	21	222	26,504	580,513
	Total	40	341	51,907	1,419,161
Test	En	7	85	-	6,466
	Zh	7	60	-	14,257
	Total	14	145	-	20,723

Table 2: Dataset statistics for B^{2}NERD. "Num" refers to the number of samples in the final B^{2}NERD. "Raw Num" represents the samples in collected datasets and B^{2}NERD$_{\text{all}}$ before data pruning.

pools. In practice, we treat the same entity type in different datasets as distinct entity types, enabling efficient per-dataset pruning. Additionally, up to $\frac{1}{5}k$ negative samples are added per dataset.

We implement the diversity-aware data pruning strategy on the training set of B^{2}NERD$_{\text{all}}$ using $k = 400$ and $b = 0$. In Section 5.4, we compare different sampling methods and data scales. The resulting B^{2}NERD dataset (Table 2), featuring a universal NER taxonomy and efficient samples, enables LLMs to generalize beyond raw dataset boundaries.

3.3 Instruction Tuning with Regularization

Based on B^{2}NERD, we conduct instruction tuning on LLMs to create the B^{2}NER models, using a UIE-style NER instruction template similar to Wang et al., 2023b (See Appendix C.2).

We observe that instructions for samples from the same dataset include a shared part about entity label options ("Label Set:[...]"), causing LLMs to mechanically memorize this part rather than understanding the actual labels. Addressing this, we introduce training regularization methods to prevent dataset-specific patterns in the instructions. A key innovation is **Dynamic Label Set**: Instead of asking LLMs to recognize a fixed set of labels for each dataset, we randomly vary the number and order of entity types mentioned in the prompts to reduce co-occurrences. See Appendix A.4 for more training regularization details.

4 Experimental Settings

4.1 Implementation

Data Our B^{2}NER models are trained on the B^{2}NERD dataset, containing 25,403 English samples and 26,504 Chinese samples. For comparison with previous works, we include Pile-NER from Zhou et al., 2024 as extra training data. Test datasets are held out for out-of-domain evaluation in Section 4.2. More statistics are in Appendix D.

Backbone We derive our models by fine-tuning InternLM2 (Cai et al., 2024) with LoRA (Hu et al., 2021). Training details are in Appendix C.1. InternLM2 is chosen for its balanced performance in English and Chinese, fitting our bilingual training data. We also validate our approach using other backbones in Appendix B.2.

4.2 Evaluation

Benchmarks As the core aspect of the Open NER task, we assess the model's out-of-domain performance on 3 benchmarks using held-out datasets from the training data. For English NER (Table 3), we follow Wang et al., 2023b and use 7 datasets from CrossNER and MIT. For Chinese NER (Table 4), we create a comprehensive OOD benchmark by holding out 7 Chinese datasets covering various domains and entity types. For multilingual NER (Table 5), we use Multiconer22 (Malmasi et al., 2022) to evaluate cross-lingual effects. These held-out datasets include both unseen and common entities, reflecting practical scenarios. Datasets with

Model	*w/ Unseen Entities* Movie	Restaurant	*w/ Common & Unseen Entities* AI	Litera.	Music	Politics	Science	Avg.	Instance/s
Non-Natural Language Prompt									
GoLLIE-7B	63.0	43.4	59.1	62.7	67.8	57.2	55.5	58.4	-
KnowCoder-7B	50.0	48.2	60.3	61.1	70.0	72.2	59.1	60.1	-
GNER-7B	68.6	47.5	63.1	68.2	75.7	69.4	69.9	66.1	4.0
GNER-11B	62.5	51.0	**68.2**	68.7	81.2	75.1	76.7	69.1	3.0
Natural Language Prompt									
InstructUIE-11B	63.0	21.0	49.0	47.2	53.2	48.2	49.3	47.3	3.4
UniNER-7B	59.4	31.2	62.6	64.0	66.6	66.3	69.8	60.0	1.6
GPT-4	60.4	**59.7**	50.0	55.2	69.2	63.4	63.2	60.1	-
Baseline-7B	49.7	36.6	43.7	44.0	58.6	59.8	60.0	50.3	-
B^{2}NER-7B (w/o English)	68.5	50.4	56.9	55.0	65.1	67.2	65.9	61.3	-
B^{2}NER-7B (only English)	67.6	53.3	59.0	63.7	68.6	67.8	72.0	64.6	-
B^{2}NER-7B	70.2	56.8	64.1	69.0	76.4	75.5	76.7	69.8	**16.1**
B^{2}NER-20B	**71.4**	57.1	64.7	**71.6**	**82.4**	**78.2**	**79.4**	**72.1**	7.0

Table 3: Out-of-domain evaluation results on English NER. **Bold** numbers highlight the best scores, while underlined numbers indicate suboptimal scores. "*w/ Unseen Entities*" denotes datasets with every entity type unseen during training. "*w/ Common & Unseen Entities*" denotes datasets with a mix of common and unseen entities. "w/o English" refers to a cross-lingual model trained without any English data. Our models use InternLM2 as the backbone LLM; results on other backbones are shown in Appendix B.2. The last column reports inference speed, tested on a single 8×A100 node with a batch size of 4 per device following Ding et al., 2024.

Model	*w/ Unseen Entities* Law (CAIL2021)	Math	Address (CCKS2021)	*w/ Common & Unseen Entities* Cluener	Medical (CBLUE)	Weibo	Onto. 4	Average (SoTA)	(All)
SoTA	-	-	68.5*	36.5†	31.4*	38.0‡	39.2§	42.7	-
GPT-4	**69.1**	45.9	70.5	55.7	44.6	34.0	68.8	54.7	55.5
Baseline-7B	52.0	44.5	65.5	55.7	43.3	33.0	73.8	54.3	52.5
B^{2}NER-7B (w/o Chinese)	58.7	60.2	56.6	51.7	43.7	38.6	70.7	52.3	54.3
B^{2}NER-7B (only Chinese)	66.6	50.9	68.4	57.1	46.1	39.5	76.3	57.5	57.9
B^{2}NER-7B	64.7	60.8	**73.0**	60.3	45.0	41.3	77.4	59.4	60.3
B^{2}NER-20B	67.6	**62.2**	71.0	**64.4**	**46.8**	**44.6**	**79.8**	**61.3**	**62.3**

Table 4: Out-of-domain evaluation results on Chinese NER. "*", "†", "‡", and "§" denote the SoTA results from Fang et al., 2023, YAYI-UIE (Xiao et al., 2023), IEPile (Gui et al., 2024), and Xie et al., 2023, respectively.

all unseen entity types, representing a stricter zero-shot evaluation, are highlighted in our result tables. We also conduct in-domain supervised evaluation on 20 English datasets from Wang et al., 2023b and 6 Chinese datasets from our collection.

Metrics Evaluation is based on strict span-based micro-F1, requiring exact entity type and boundary matching. Experiments are repeated four times, and the results are averaged.

4.3 Compared Systems

For English NER, we primarily compare our model with **InstructUIE** (Wang et al., 2023b) and **UniversalNER** (Zhou et al., 2024), which use similar training data and natural language prompts to us. We also include strong generative NER systems that don't use natural language prompts, such as code-based **GoLLIE** (Sainz et al., 2024), **KnowCoder** (Li et al., 2024) and BIO tag-based **GNER** (Ding et al., 2024). Additionally, we train a **Baseline** model with the same data sources and backbone as our B^{2}NER models but without dataset normalization approaches.

For Chinese NER, we compare with SoTA zero-shot or OOD NER systems including Xie et al., 2023, **YAYI-UIE** (Xiao et al., 2023) and **IEPile** (Gui et al., 2024). We also include 1-shot NER results from Fang et al., 2023.

Moreover, we compare our models with **GPT-4** (Achiam et al., 2023), a milestone proprietary LLM. For both English and Chinese, we prompt GPT-4-0613 to perform entity recognition on test datasets using the same instructions and standardized label names as our model. To ensure fairness, we fix format issues in GPT-4's responses.

5 Experiment Results

5.1 Out-of-domain Evaluation

Comprehensive experiments across languages and datasets demonstrate our method's effectiveness in improving out-of-domain generalization.

English NER Table 3 shows the out-of-domain evaluation results on the English NER benchmark. Both B^{2}NER-7B and B^{2}NER-20B exhibit superior average performance over previous methods and surpass GPT-4 by 9.7–12.0 F1 points, demonstrating their advanced capabilities. Compared to Baseline, InstructUIE, and UniNER, which use similar data sources and prompts, B^{2}NER-7B significantly improves for all 7 datasets with unseen or mixed entities, highlighting the value of our normalized data. Moreover, B^{2}NER-7B (69.8%) slightly surpasses the previous SoTA GNER-11B (69.1%) despite its smaller size and achieves a much faster inference speed (4X) than GNER-7B. This speed stems from our generic UIE-style prompt (See Appendix C.2) that extracts only relevant content, unlike GNER's prompts that generate all text with tags, leading to longer responses and less flexibility. Additionally, we observe a surprising cross-lingual effect: our "w/o English" model trained without any English data achieves comparable performance to GPT-4 on English, showing that the learned universal taxonomy can transfer between languages.

Chinese NER Table 4 presents the out-of-domain evaluation results on our Chinese NER benchmark. Both our 7B and 20B models outperform GPT-4 and other methods, exceeding the previous SoTA by 18.6 points on average. B^{2}NER-7B substantially improves upon the Baseline model for all 7 datasets with unseen or mixed entities, further validating the value of our normalized data. Moreover, B^{2}NER-7B boosts the average performance of the "only Chinese" model on Chinese and the "only English" model on English, showing that joint training with our bilingual NER data enhances performance in both languages. This suggests our universal taxonomy addresses the data disparity concerns of bilingual training, as discussed by Gui et al., 2024.

Multilingual NER Table 5 shows the out-of-domain evaluation results on the multilingual dataset `Multiconer22`. We include 6 languages that constitute more than 0.1% of the general LLM pretraining corpus (Touvron et al., 2023). For strict OOD evaluations, we exclude all `Multiconer22` and `Multiconer23` samples from our training data. We compare our model with ChatGPT (evaluated by Lai et al., 2023) and GLiNER (Zaratiana et al., 2023), which uses mdeBERTa-v3-base as backbone. From the results, our model achieves the best performance on 5 out of 6 languages. In the cross-lingual setting, without any training data in the target languages, our method improves the baseline model from 28.2% to 39.9%, outperforming other unsupervised methods and showing that learning a universal taxonomy benefits LLM generalization across language boundaries.

Language	Sup.	ChatGPT	GLiNER	Base.	Ours-7B
English	62.7	37.2	41.7	39.8	**54.8**
Chinese	53.1	18.8	24.3	32.8	**45.4**
Cross-Lingual					
German	64.6	37.1	**39.5**	26.5	36.6
Spanish	58.7	34.7	42.1	34.1	**46.0**
Dutch	62.6	35.7	38.9	32.2	**43.0**
Russian	59.7	27.4	33.3	19.9	**33.9**
$\text{Average}_{\text{cross}}$	61.4	33.7	38.5	28.2	**39.9**
$\text{Average}_{\text{all}}$	60.2	31.8	36.6	30.9	**43.3**

Table 5: Out-of-domain multilingual evaluation results on `multiconer22`. "Sup." indicates supervised baseline results from Malmasi et al., 2022. "Base." denotes the baseline model trained without dataset normalization.

Model	EN Avg. (20 Datasets)	ZH Avg. (6 Datasets)	Avg.
BERT-based	80.09	84.74	82.42
InstructUIE-11B	81.16	-	-
UniNER-7B	**84.78**	-	-
B^{2}NER-7B	83.85	**85.11**	**84.48**

Table 6: In-domain supervised evaluation results on 20 English and 6 Chinese datasets. "EN" and "ZH" denote results on English and Chinese, respectively. The full table with details can be found in Appendix B.1.

5.2 In-domain Supervised Evaluation

Despite our focus on out-of-domain generalization, we also conduct in-domain supervised experiments. In this setting, we train and evaluate our B^{2}NER model on 20 English datasets (Wang et al., 2023b) and 6 Chinese datasets from our $\mathtt{B^2NERD_{all}}$ collection with standardized entity labels. Following previous work, we sample 10,000 examples for each dataset instead of using our pruning strategy. The training arguments slightly differ from those used in OOD experiments (see Appendix C.1). We compare our model with BERT-based task-specific models, using English results from Wang et al., 2023b and Chinese results from our evaluation.

Model	EN	ZH	OOD Avg.
B²NER-7B	69.8	60.3	65.1
w/o entity definition std.	62.4	58.5	$60.5_{\downarrow 4.6}$
w/ dataset names	60.5	57.0	$58.8_{\downarrow 6.3}$
w/o data pruning	66.2	56.6	$61.4_{\downarrow 3.7}$
w/o training regularization	68.5	59.7	$64.1_{\downarrow 1.0}$
w/o *all above* (Baseline)	50.3	52.5	$51.4_{\downarrow 13.7}$
w/o `Pile-NER`	69.5	60.2	$64.9_{\downarrow 0.2}$
GPT-4	60.1	55.5	57.8
w/o entity definition std.	53.0	50.6	$51.8_{\downarrow 6.0}$

Table 7: Ablation study for both B²NER and GPT-4. F1 scores come from out-of-domain evaluations.

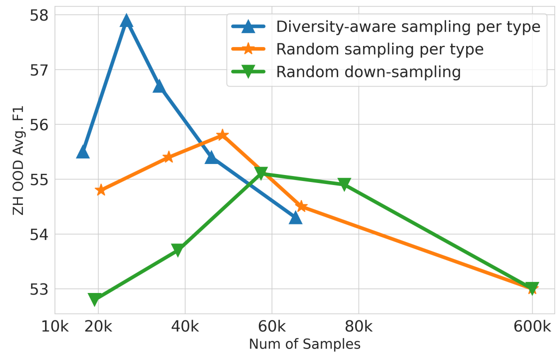


Figure 5: Data scaling results for different sampling methods. Diversity-aware strategy (blue) achieves better performance with fewer samples.

Results are shown in Table 6. B²NER-7B achieves better average performance than BERT-based models on both English and Chinese datasets. For the 20 English datasets, B²NER-7B outperforms InstructUIE-11B and slightly trails UniNER-7B by 1 point. These results demonstrate that our approach holistically enhances Open NER capabilities, achieving both superior out-of-domain generalization and competitive in-domain performance.

5.3 Ablation Study

Table 7 details our ablation study on the impact of various components in our approach under OOD evaluations. Results on B²NER show significant benefits from entity definition standardization, diversity-aware data pruning, and training regularization. In contrast, the "w/o `Pile-NER`" model trained solely with B²NERD data, shows minimal performance regression, indicating the individual effectiveness for our data. Additionally, adding dataset names ("w/ dataset names") instead of standardizing entity definitions hurts overall performance, confirming that models learn dataset-specific patterns this way (See Appendix B.5 for a case study). For GPT-4, skipping the entity definition standardization on test datasets also leads to substantial performance losses, underscoring the overall effectiveness of our entity definition standardization and universal taxonomy on LLMs.

5.4 In-depth Analysis of Data Pruning

To better understand the impact of our diversity-aware data pruning method, we compare various sampling strategies and examine data scaling effects. We focus on out-of-domain setting and experiment on Chinese NER data for simplicity.

Sampling Strategies Beyond our diversity-aware strategy, we evaluate 2 additional methods: 1) Random sampling per type, which evenly selects random samples for each entity type, ignoring the semantic diversity. 2) Random down-sampling, which selects random samples regardless of entity types. Figure 5 shows that diversity-aware sampling achieves the highest peak performance, while random down-sampling yields the weakest. This highlights the importance of data diversity in tuning LLMs. Variants of diversity-aware sampling strategies are explored in Appendix B.4.

Data Scaling We varied the value of k during data pruning to generate datasets of differing scales for our experiments. The line plots in Figure 5 show that for all sampling strategies, peak performance is achieved with moderate data size; both excess and scarcity of data can hinder model effectiveness. Another clear trend is that **diversity-aware strategy can achieve better performance with fewer samples**. These results support our assumption that redundant data causes LLMs to over-fit, while curated, diverse data benefits universal generalization. This finding aligns with recent data efficiency research (Zhou et al., 2023; Liu et al., 2024; Ye et al., 2024).

6 Discussion

Error Analysis Despite achieving state-of-the-art generalization by training on our compact and coherent dataset, our best model still produces many errors in current out-of-domain evaluations, with top performances of 72.1 on English and 62.3 on Chinese benchmarks. To guide future work, we analyzed some error cases and identified a key issue: the model struggles to align with the unique annotation standards of each test dataset. For example, Table 8 presents representative errors from our analysis of the `Restaurant` dataset, where model predictions are reasonable but do not align with

Sentence	Prediction	Truth
any restaurants open right now	(OPERATING HOURS: **right row**)	(OPERATING HOURS: **open right row**)
can i see hamburger restaurants nearby	(**CUISINE TYPE**: hamburger)	(**DISH OR BEVERAGE NAME**: hamburger)

Table 8: Representative errors from the OOD evaluation on the `Restaurant` dataset. **Bold** text highlights the errors. While the model's predictions are reasonable, they do not align with the dataset's specific annotation conventions.

the dataset's specific conventions. These minor, unique inconsistencies are hidden and pervasive across test datasets, making them difficult to address through written descriptions and challenging for out-of-domain models to capture without extensive fine-tuning. We view this type of error as a limitation of current evaluation benchmarks and methods. Therefore, a promising direction for future work is to develop a dedicated benchmark or evaluation method for more accurate assessment of strong Open NER models.

Flat and Nested NER Our method supports both flat NER and nested NER tasks, but current dataset is mainly developed and tested for flat NER tasks. There are 2 nested NER datasets in our collected datasets: `ACE2005` and `GENIA`. We assign distinct entity labels to nested datasets during our standardization process to prevent conflicts with flat datasets, so current models will only extract nested entities for those specific labels. The dataset can be easily reused or extended to train models with better generalization for nested NER tasks by incorporating explicit hints in prompts (e.g., a "nested" tag) or entity labels for nested NER training data.

7 Related Work

Instruction Tuning Instruction tuning (Sanh et al., 2022; Ouyang et al., 2022) can boost LLMs' efficacy on unseen tasks via fine-tuning with exemplary natural language instructions. Current instruction tuning datasets, constructed from human (Conover et al., 2023), LLM (Wang et al., 2022a; Xu et al., 2023a), or existing datasets (Longpre et al., 2023; Wang et al., 2022b; Yu et al., 2023), mostly prioritize large quantities. In contrast, recent work (Zhou et al., 2023; Liu et al., 2024; Ye et al., 2024) shows that using fewer but higher quality instruction tuning data could align LLMs better on general tasks. Our work, following this direction, focuses on downstream applications like Open NER, where data engineering strategies on how to merge and prune task-specific datasets for efficient instruction tuning are still under-explored.

Generative NER Numerous attempts have been made to harness LLMs to solve Information Extraction (IE) tasks like NER in a generative paradigm (Xu et al., 2023b). Researchers (Xie et al., 2023; Wang et al., 2023a; Ashok and Lipton, 2023) leverage LLMs like ChatGPT for NER via in-context learning, which is orthogonal to our approach (see Appendix B.3). Recent studies use instruction tuning to train custom LLMs with existing datasets (Wang et al., 2023b; Gui et al., 2024; Xiao et al., 2023), but face challenges in Open NER due to dataset inconsistencies and redundancies. GoLLIE (Sainz et al., 2024) trains LLMs to follow detailed code-style annotation guidelines to resolve inconsistent entity definitions, but such guidelines can be difficult to obtain and understand. Our work takes a different approach that directly clarifies entity ambiguities and restructures existing datasets for optimal LLM learning. Other studies explore synthetic NER data distilled from LLMs (Zhou et al., 2024; Lu et al., 2023; Li et al., 2024; Ding et al., 2024). However, synthetic data often falls short in covering real-world NER tasks comprehensively.

8 Conclusion

We present $\mathtt{B^2NERD}$, a cohesive and compact dataset designed to enhance LLMs for Open NER. Refined from 54 datasets through entity definition standardization and diversity-aware data pruning, $\mathtt{B^2NERD}$ addresses inconsistencies and redundancies in existing datasets, enabling LLMs to learn a universal entity taxonomy beyond data boundaries. Models trained on $\mathtt{B^2NERD}$ outperform GPT-4 and previous methods in out-of-domain evaluations across various datasets and languages. We will share our recipe and data to support further research.

Limitations

While our work contributes to stronger LLMs for the Open NER task, it has following limitations:

- **Benchmarks**: Current out-of-domain evaluation is mainly performed by holding out certain datasets from existing ones. However, these test datasets may contain unique annotation standards that can't be learned via OOD generalization and may suffer from data contamination. Based on our error analysis for our 20B model, we are concerned that the ceiling for these datasets' OOD evaluation may soon be approached. A dedicated comprehensive benchmark for Open NER evaluation may be necessary in the near future.

- **Diversity Measure**: In our existing data pruning strategy, we evaluate the diversity of entity types and text semantics independently. Semantic diversity is assessed within the context of each entity type. Yet, a more inclusive measure could be developed to simultaneously compare annotations and text in pairs of samples. Such an approach might enable globally optimal data selection, encapsulating more information with fewer samples and providing insights on what kind of data are best for task-focused instruction tuning.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 62376061, 62206057, 62076069), the Major Key Project of PCL under Grant PCL2024A06, Shanghai Rising-Star Program (23QA1400200), Natural Science Foundation of Shanghai (23ZR1403500), Program of Shanghai Academic Research Leader under grant 22XD1401100.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30- May 2, 2015*.

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. 2022. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3329–3339.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179.

Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Bowen Yan, and Min Zhang. 2024. Rethinking negative instances for generative named entity recognition. *arXiv preprint arXiv:2402.16602*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Jinyuan Fang, Xiaobin Wang, Zaiqiao Meng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Manner: A variational memory-augmented model for cross domain few-shot named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4261–4276.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2). *arXiv preprint arXiv:2305.06586*.

Jun Gao, Huan Zhao, Yice Zhang, Wei Wang, Changlong Yu, and Ruifeng Xu. 2023. Benchmarking large language models with augmented instructions for fine-grained information extraction. *arXiv preprint arXiv:2310.05092*.

Runwei Guan, Ka Lok Man, Feifan Chen, Shanliang Yao, Rongsheng Hu, Xiaohui Zhu, Jeremy Smith, Eng Gee Lim, and Yutao Yue. 2024. Findvehicle and vehiclefinder: a ner dataset for natural language-based vehicle retrieval and a keyword-based cross-modal vehicle retrieval system. *Multimedia Tools and Applications*, 83(8):24841–24874.

Honghao Gui, Hongbin Ye, Lin Yuan, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024. Iepile: Unearthing large-scale schema-based information extraction corpus. *arXiv preprint arXiv:2402.14710*.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

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*.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of NAACL*.

Uri Katz, Matan Vetzler, Amir Cohen, and Yoav Goldberg. 2023. Neretrieve: Dataset for next generation named entity recognition and retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3340–3354.

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.

Veysel Kocaman and David Talby. 2021. Biomedical named entity recognition at scale. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, pages 635–646. Springer.

Aman Kumar and Binil Starly. 2022. "fabner": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*, 33(8):2393–2407.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *ArXiv*, abs/2304.05613.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN workshop on Chinese language processing*, pages 108–117.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xiang Li, Zhilei Hu, et al. 2024. Knowcoder: Coding structured knowledge into llms for universal information extraction. *arXiv preprint arXiv:2403.07969*.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. Pivoine: Instruction tuning for open-world information extraction. *arXiv preprint arXiv:2305.14898*.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 548–554.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. Multinerd: 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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Asahi Ushio, Leonardo Neves, Vitor Silva, Francesco Barbieri, and Jose Camacho-Collados. 2022. Named entity recognition in twitter: A dataset and analysis on short-term temporal shifts. *arXiv preprint arXiv:2210.03797*.

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

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. LDC2011T03.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2023. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction. *arXiv preprint arXiv:2312.15548*.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot ner with chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023b. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

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*.

Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *ArXiv*, abs/2303.10420.

Junjie Ye, Yuming Yang, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2024. Empirical insights on fine-tuning large language models for question-answering. *arXiv preprint arXiv:2409.15825*.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, et al. 2023. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. *arXiv preprint arXiv:2308.10529*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer. *ArXiv*, abs/2311.08526.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022a. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2022b. Optimizing bi-encoder for named entity recognition via contrastive learning. *arXiv preprint arXiv:2208.14565*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Implementation Details

A.1 Details of Data Collection and Cleaning

We spend non-trivial effort on data collection and data cleaning for Chinese NER data by following major steps. **Data collection**, after an extensive search, we initially identify 35 publicly available Chinese NER datasets, about half of which is never used by previous works. **Deduplication**, We remove those datasets that have highly duplicate data with others. For example, `peopledaily1998` dataset is actually part of `MSRA` dataset. **Annotation quality screening**, as many datasets didn't share details on their labelling process, we manually re-evaluate their annotation consistency at dataset level. Datasets with low internal consistency are excluded. **Label name translation**, many datasets use English symbols and Arabic numbers as entity type name, such as "PER", "HCCX", "1". To help LLM understand the entity type together with input Chinese text, we translate all type names into natural language in Chinese. We prompt GPT4 to help the translation. These label names will be further standardized in Section 3.1.

A.2 Details of Automatic Dataset Conflict Detection

For model-based cross validation, we implement a BERT-CRF model[2] to learn from one dataset and infer on others with similar entity types. Figure 6 shows more results from this cross validation.

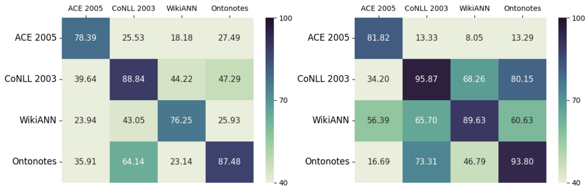
To better understand the conflicts reflected in model-based cross-validation results, we examine the surprisingly low F1 scores for the LOC entity across two popular datasets, `CoNLL` and `OntoNotes` (Figure 2). Our analysis indicates that these discrepancies indeed arise from differences in entity definitions. `OntoNotes` defines LOC as "non-GPE locations, mountain ranges, and bodies of water" with a separate GPE type for geo-political entities[3], while `CoNLL` includes GPE within LOC. Additionally, many LOC mentions in `CoNLL`, such as "New Zealand" and "Minnesota," are actually GPE entities, due to the dataset's news source. Thus, as shown in Table 9, the model trained on `CoNLL` tends to mislabel `OntoNotes`'s GPE entities as LOC, resulting in low accuracy. Conversely, the reverse model annotates many LOC entities in `CoNLL` as GPE, leading to low recall.

[2] https://github.com/lonePatient/BERT-NER-Pytorch

[3] https://catalog.ldc.upenn.edu/docs/LDC2011T03/OntoNotes-Release-4.0.pdf

Train → Test	Precision	Recall	F1
CoNLL → OntoNotes	4.64	65.05	8.66
OntoNotes → CoNLL	64.52	1.20	2.35

Table 9: Detailed metrics for the model-based cross validation between `CoNLL` and `OntoNotes` for LOC entity.



Cross-dataset validation for "**ORG**" Cross-dataset validation for "**PER**"

Figure 6: More results from model-based cross validation on PER and ORG among 4 datasets. The horizontal axis represents testing data and vertical represents training data.

For rule-based screening, we employ detailed rules to ensure accurate detection. We screen for cases where 1) two datasets share one entity label (e.g., LOCATION); 2) the same entity mention appeared in the samples of both datasets (e.g., "Belgium"); 3) this mention receives inconsistent recognition results in two datasets (e.g., LOCATION v.s. GEO-POLITICAL ENTITY). We also exclude cases where current mention is part of another extracted entity, as it's reasonable for flat NER datasets not to extract this mention. After screening, we classify the error types of inconsistent cases for each pair of conflicting entities and list significant ones for future processing. Error types include wrong categories, not extracted and partially extracted.

A.3 Details of Entity Definition Standardization

We publicly recruit 4 college students with prior NER annotation experience as human experts to delineate entity definition differences, re-assign label names, and write annotation guidelines (for the experiments in Appendix B.3). These experts, including one with a biomedical background, are selected for their strong understanding of entity definitions and annotation guidelines. Since the task involves providing proper entity label names for existing entity types on a dataset basis rather than re-labeling individual samples, their prior experience and knowledge enable them to perform the task with sufficient expertise after some training. Compensation for the annotators is provided on an hourly basis and exceeds local standards, reflecting the effort and expertise required for this task.

Provided Resources and Instructions The annotators are equipped with several resources to aid their work, including conflict detection results (as described in Appendix A.2), which show conflict statistics, error types of conflicting entity definitions, and examples of the conflicts. They also receive raw datasets, sample annotations from these datasets, and data summaries, such as the most frequent entity mentions. Additionally, our naming principles with examples (outlined in Section 3.1) and supplementary materials, such as the ACE annotation guidelines, are provided. Based on this information, annotators are tasked with screening each entity type, writing proper label names, and justifying their choices based on the naming principles and previous annotation guidelines.

LLMs as an Auxillary Tool LLMs are also used to accelerate the process. Annotators receive a standard prompt instructing the LLM to suggest proper label names based on entity annotation examples and our naming principles. While annotators are welcome to use their own prompts to explore additional insights, the final label names are determined through careful human consideration. The LLM suggestions primarily act as a reference, ensuring that human expertise remains central to the decision-making process.

Time Allocation and Workflow Each expert undergoes 3 hours of initial training to familiarize themselves with the task, resources, and naming principles. They then spend 12 hours screening and renaming approximately 120 assigned entity types each, based on detected conflicts and guidelines. Following this, 5 hours are allocated for group discussions, during which the experts collectively review and check the consistency of all entity labels, finalize names, and resolve any ambiguities. In total, the four annotators collectively spend approximately 80 hours on this process.

Attention Checks During the final discussion phase, the authors review the annotators' work for quality by examining the provided entity labels and the justifications for each. Ill-defined entity types are also eliminated as part of this process. Common entities encountered by most annotators serve as attention checks to ensure accuracy and consistency. The high-quality work provided by the annotators, combined with the rigorous group discussions, ensures that the final universal taxonomy is both accurate and consistent.

Dataset	BERT-base	InstructUIE-11B	UniNER-7B	B^{2}NER-7B
ACE05	**87.30**	79.94	86.69	83.04
AnatEM	85.82	88.52	88.65	**89.18**
bc2gm	80.90	80.69	**82.42**	81.95
bc4chemd	86.72	87.62	**89.21**	88.96
bc5cdr	85.28	89.02	**89.34**	88.52
Broad Twitter	58.61	80.27	81.25	**82.16**
CoNLL03	92.40	91.53	**93.30**	92.56
FabNER	64.20	78.38	**81.87**	78.82
FindVehicle	87.13	87.56	**98.30**	97.89
GENIA	73.3	75.71	**77.54**	76.43
HarveyNER	**82.26**	74.69	74.21	73.67
MIT Movie	88.78	89.58	90.17	**90.78**
MIT Restaurant	81.02	82.59	82.35	**83.71**
MultiNERD	91.25	90.26	93.73	**93.98**
ncbi	80.20	86.21	**86.96**	84.83
OntoNotes	**91.11**	88.64	89.91	84.31
PolyglotNER	**75.65**	53.31	65.67	61.96
TweetNER7	56.49	65.95	65.77	**66.26**
WikiANN	70.60	64.47	84.91	**85.07**
wikiNeural	82.78	88.27	**93.28**	93.01
Avg	80.09	81.16	**84.78**	83.85

Table 10: Full results of in-domain supervised evaluation on English NER.

Dataset	BERT-base	YAYI-UIE	IEPile	B^{2}NER-7B
CCKS2017	92.68	90.73	-	**94.93**
MSRA	**96.72**	95.57	87.99	92.22
Multiconer22	69.78	-	-	**71.53**
Multiconer23	66.98	-	-	**69.56**
resume	**96.01**	-	93.92	95.90
Youku	86.26	-	-	**86.50**
Avg	84.74	-	-	**85.11**

Table 11: Full results of in-domain supervised evaluation on Chinese NER.

A.4 Details of Training Regularization

We observe that the co-occurrence of entity label options in instructions is an obvious pattern for samples coming from same original dataset. For example, data from `Taobao` dataset all ask to recognize PRODUCT NAME and BRAND in their instructions. LLMs may just memorize this dataset-specific co-occurrence pattern without understanding the given label names. To alleviate this, we introduce regularization methods, including: **Dynamic label set.** Instead of asking LLM to recognize a static set of labels, we mention random entity types in random order in instructions for less co-occurrence patterns. Labels that current sample contains still remain in instructions to assure the answer is still correct. **Random label dropout.** We randomly neglect some entity types in both the instruction and answer of a sample. This can force LLM to focus on target label names in instructions when generating answers.

Model	EN	ZH	OOD Avg.
B²NER-InternLM2-7B	**69.8**	**60.3**	**65.1**
B²NER-Baichuan2-7B	67.9	57.9	62.9
B²NER-Llama2-7B	66.7	42.6	54.7
GPT-4	60.1	55.5	57.8

Table 12: Out-of-domain evaluation results of fine-tuning different backbone models with our B²NERD dataset.

Model	0-shot	w/ guidelines	3-shot
Baseline (only Chinese)	52.5	56.5	58.4
B²NER (only Chinese)	57.9	62.3	**62.6**

Table 13: Study on additional in-context learning methods with annotation guidelines or few-shot examples. "0-shot" denotes our out-of-domain evaluation using zero-shot instruction template.

B More Experiments and Studies

B.1 In-domain Supervised Evaluation Results

Table 10 shows the full results for in-domain supervised evaluation on English NER. Though trailing UniNER-7B by 1 point on average, B²NER-7B achieves best results in 7 out of 20 datasets.

Table 11 shows the full results for in-domain supervised evaluation on Chinese NER. We do not use the complete Chinese training datasets for in-domain supervised evaluation because some datasets lack high-quality test sets in their original splits. Our B²NER-7B model achieves the best performance on 4 out of 6 datasets and surpasses BERT-based models on average.

B.2 Results of Different Backbones

We investigate the effectiveness of our approach and the B²NERD dataset on different backbone models. In addition to InternLM2-7B, we further fine-tune our dataset on Baichuan2-7B (Baichuan, 2023) and Llama2-7B (Touvron et al., 2023).

As shown in Table 12, all models achieve superior out-of-domain performance over GPT-4 except B²NER-Llama2-7B, which trails behind on Chinese NER. B²NER-InternLM2-7B achieves best overall performance.

B.3 Compatibility with In-Context Learning

As our method is orthogonal to other in-context learning approaches, such as adding annotation guidelines (Sainz et al., 2024) and few-shot examples (Wang et al., 2023a), we explore their combined performance. Focusing on Chinese NER, we invite experts to write guidelines for Chinese datasets in B²NERD. Models are trained with these guidelines or few-shot examples using instruction templates in Appendix C.2.

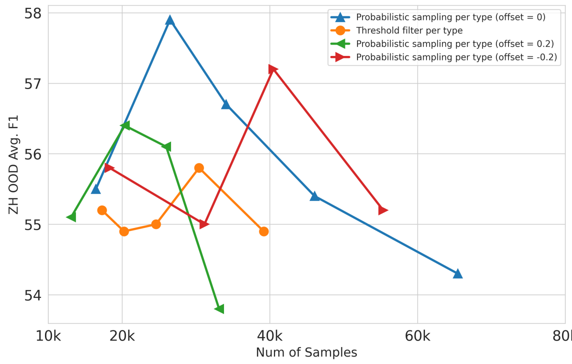


Figure 7: Data scaling results for variants of diversity-aware sampling strategy. Higher offset means more strict semantic distance requirement.

Results in Table 13 show that our "B²NER (only Chinese)" model can be further improved with in-context learning, demonstrating the compatibility. Notably, the Baseline model with guidelines still fails to outperform B²NER (56.5 < 57.9), highlighting our approach's superior effectiveness over additional guidelines.

B.4 Variations of Diversity-aware Sampling Strategy

We experimented with other diversity-aware sampling strategies during data pruning. One alternative is the "threshold filter per type," which uses a hard semantic distance threshold instead of probabilistic sampling for selecting samples for each entity type's pool. We also tried different offsets b for the semantic distance measure, as introduced in Section 3.2. A higher offset imposes a stricter semantic distance requirement, resulting in more diverse semantics.

Figure 7 shows that an offset of 0 achieves the best peak performance. Additionally, higher offsets reach peak performance with fewer samples, indicating that greater semantic diversity can compress information into a smaller dataset.

B.5 Case Study on Dataset-Specific v.s. General Patterns

Figure 8 shows an out-of-domain NER example on CCKS2021address dataset. The baseline model trained with raw inconsistent datasets produces incorrect and out-of-scope entity types, reflecting its learning of dataset-specific patterns from prior City-Location style data. In contrast, Our B²NER

Instruction:
请识别文本中属于[城市, 区县, 兴趣点]的所有实体
(Please identify all entities of [City, District, Points of Interest] in the text.)

Text:
路桥区方林汽车城
(Fanglin Automobile Center, Luqiao District)

Answer:

Baseline Model 城市: 路桥区; 地名: 方林汽车城
(City: Luqiao District; Location: Fanglin Automobile Center)

***B*²NER** 区县: 路桥区; 兴趣点: 方林汽车城
(District: Luqiao District; Points of Interest: Fanglin Automobile Center)

Figure 8: An out-of-domain NER example on the dataset of `CCKS2021address`, where baseline model displays dataset-specific patterns.

delivers accurate results for this unseen task, owing to our instruction tuning method that transcends data boundaries.

B.6 Experiments on Two-stage Training

Previous studies, such as Ding et al., 2024, use a two-stage training strategy, starting with general `Pile-NER` data followed by supervised datasets. We test this approach on the OOD English NER benchmark.

	Two-Stage Training	Mixed Training
OOD Avg. F1	69.7	69.8

Table 14: Comparison between two-stage training and mixed training strategies

In Table 14, "Two-Stage Training" involves sequential training with `Pile-NER` and B^2`NERD`, while 'Mixed Training' refers to our strategy of training them simultaneously. The similar performance of both approaches suggests that each is equally effective in our LoRA training scenario.

C Training Details

C.1 Hyper-parameters

As explained in Section 4.1, we use InternLM2 (Cai et al., 2024) as the backbone model and fine-tune it with LoRA (Hu et al., 2021) to derive all the models. Although we also experiment with other bilingual backbone LLMs in Appendix B.2, InternLM2 demonstrates better overall performance.

For our main out-of-domain experiments, we apply LoRA to "wqkv" target modules, setting r to 32 and the dropout rate to 0.05. Preliminary comparisons between full-parameter tuning and LoRA show that LoRA provides better and more stable results. During training, we use 3e-4 learning rate with warmup ratio of 0.02 and a cosine scheduler. DeepSpeed (Rasley et al., 2020) Stage 2 is adopted for memory optimization. The main model, B^{2}NER, is trained with a batch size of 128. For datasets of varying sizes, we experimented with different batch sizes to find the most effective configuration. The maximum context length of our LLM is set to 4096 tokens. Training and inference are done on one 8$\times$ Nvidia-A100-40G node, with a single run of 5 epochs taking about 8 hours.

For in-domain supervised experiments, we use full-parameter tuning with a learning rate of 2e-5 and disable the training regularization methods in Appendix A.4. All other settings match the out-of-domain experiments.

Instruction: Given the label set of entities, please recognize all the entities in the text. The answer format should be \"entity label: entity; entity label: entity\".

Label Set: [Type A, Type B, …]

Text: [Text]

Answer: [Answer]

Figure 9: Instruction template for our main experiments.

Instruction: Given the label set of entities, please recognize all the entities in the text. The answer format should be \"entity label: entity; entity label: entity\".

Label Set: [Type A, Type B, …]

Guidelines: A means [guideline A], B means…

Text: [Text]

Answer: [Answer]

Figure 10: Instruction template with annotation guidelines for experiments in Appendix B.3.

Instruction: Given the label set of entities, please recognize all the entities in the text. The answer format should be \"entity label: entity; entity label: entity\". Below are some examples.

Label Set: [Type A, Type B, …]

Text: [Example1 text]

Answer: [Example1 answer]

Text: [Example2 text]

Answer: [Example2 answer]

…

Text: [Text]

Answer: [Answer]

Figure 11: Few-shot instruction template for experiments in Appendix B.3.

C.2 Instruction Templates

Figure 9 displays the instruction template used in our main experiments. For the study in Appendix B.3, Figure 10 shows our instruction template when annotation guidelines are available; Figure 11 shows the template for our few-shot experiments.

D Dataset and Taxonomy

D.1 Dataset Statistics

Table 15 shows the statistics of all English datasets inside B^{2}NERD. Table 16 shows the statistics of all Chinese datasets.

All datasets are flat NER datasets, except for two nested NER datasets: ACE2005 and GENIA. Note that for CAIL2021, we randomly split 20% samples as test set for evaluation. Other datasets all inherit original train and test splits.

D.2 Full NER Taxonomy

Table 17 shows the full NER taxonomy of English entities in B^{2}NERD. Table 18 shows the full NER taxonomy of Chinese entities.

Dataset	Types	Major Domain	Train	Test	Pruned Train Num	Raw Train Num	Test Num
ACE2004(Mitchell et al., 2005)	7	News	✓		1193	6177	812
ACE2005(Walker et al., 2006)	7	News	✓	✓	1433	7134	1050
AnatEM(Pyysalo and Ananiadou, 2014)	1	Biomedical	✓	✓	480	5667	3758
bc2gm(Kocaman and Talby, 2021)	1	Biomedical	✓	✓	480	12392	4977
bc4chemd(Kocaman and Talby, 2021)	1	Biomedical	✓	✓	480	30488	26204
bc5cdr(Zhang et al., 2022b)	2	Biomedical	✓	✓	592	4545	4788
Broad Tweet Corpus(Derczynski et al., 2016)	3	Social Media	✓	✓	855	5324	2000
CoNLL 2003(Sang and De Meulder, 2003)	4	News	✓	✓	1069	12613	3184
FabNER(Kumar and Starly, 2022)	12	Scientific	✓	✓	1741	9421	2064
FindVehicle(Guan et al., 2024)	21	Traffic	✓	✓	2591	21547	20769
GENIA_NER(Kim et al., 2003)	5	Biomedical	✓	✓	1281	14966	1850
HarveyNER(Chen et al., 2022)	4	Social Media	✓	✓	405	3553	1260
MultiNERD(Tedeschi and Navigli, 2022)	16	Wikipedia	✓	✓	4659	130623	9994
ncbi(Doğan et al., 2014)	1	Biomedical	✓	✓	480	5432	940
Ontonotes(Hovy et al., 2006)	18	General	✓	✓	4343	54994	7782
PolygoltNER(Al-Rfou et al., 2015)	3	Wikipedia	✓	✓	0	393941	10000
TweetNER7(Ushio et al., 2022)	7	Social Media	✓	✓	1325	7111	576
Wikiann en(Rahimi et al., 2019)	3	Wikipedia	✓	✓	856	20000	10000
WikiNeutral(Tedeschi et al., 2021)	3	Wikipedia	✓	✓	1140	92720	11597
CrossNER_AI(Liu et al., 2021)	14	AI		✓	/	/	431
CrossNER_literature(Liu et al., 2021)	12	Literary		✓	/	/	416
CrossNER_music(Liu et al., 2021)	13	Musical		✓	/	/	465
CrossNER_politics(Liu et al., 2021)	9	Political		✓	/	/	650
CrossNER_science(Liu et al., 2021)	17	Scientific		✓	/	/	543
mit-movie(Liu et al., 2013)	12	Social Media		✓	/	9707	2441
mit-restaurant(Liu et al., 2013)	8	Social Media		✓	/	7658	1520

Table 15: Statistics of English datasets in B^{2}NERD.

Dataset	Types	Major Domain	Source	Train	Test	Pruned Train Num	Raw Train Num	Test Num
Bank[a]	4	Social Media	Online Forum	✓		1224	10000	/
Boson[b]	6	News	News	✓		876	2000	/
CCKS2017[c]	5	Biomedical	Medical Records	✓	✓	505	2006	223
CCKS2018[d]	5	Biomedical	Medical Records	✓		212	797	/
CCKS2019[e]	6	Biomedical	Medical Records	✓		422	1379	/
CCKS2020[f]	6	Biomedical	Medical Records	✓		461	1450	/
Chinese Literature(Xu et al., 2017)	10	Literature	Literature	✓		1675	24165	2837
Finance-sina	4	News	News	✓		664	1579	/
GAIIC2022[g]	50	Social Media	Product Titles	✓		2061	6776	/
IMCS21[h]	5	Biomedical	Medical Conversations	✓		1773	98529	32935
MSRA(Levow, 2006)	3	News	News	✓	✓	867	45000	3442
Multiconer22(Malmasi et al., 2022)	6	Mixed	Wikipedia+Question+Queries	✓	✓	1889	15300	151661
Multiconer23(Fetahu et al., 2023)	33	Mixed	Wikipedia+Question+Queries	✓	✓	4839	9759	20265
NLPCC2018[i]	15	Voice	Voice Assistants	✓		1754	21352	/
Peopledaily2014	4	News	News	✓		1084	286268	/
Resume(Zhang and Yang, 2018)	8	News	Resume	✓	✓	986	3821	477
Ruijinmcc diabetes[j]	15	Biomedical	Medical Books + Papers	✓		2680	24157	2682
Taobao(Jie et al., 2019)	4	Social Media	Product Titles	✓		982	6000	1000
Wanchuang[k]	13	Biomedical	Drug Description	✓		506	1255	/
Microbiota[l]	7	Biomedical	Medical News	✓		71	99	/
Youku(Jie et al., 2019)	3	Social Media	Video Titles	✓	✓	972	8001	1001
FNED[m]	7	News	News			0	10500	/
Military-ner[n]	3	Military	Military			0	320	80
CAIL2021[o]	10	Law	Case description		✓	/	4197	1050
CCKS2021address[p]	17	Address	Address		✓	/	8856	1970
CLUENER(Xu et al., 2020)	10	News	News		✓	/	10748	1343
CBLUE(Zhang et al., 2022a)	9	Biomedical	Medical books		✓	/	15000	4999
Math-high[q]	2	Math	Math books		✓	/	1953	279
Weibo(Peng and Dredze, 2015)	8	Social media	Social media		✓	/	1350	270
Zh-Ontonotes(Weischedel et al., 2011)	4	News	News		✓	/	15724	4346

Table 16: Statistics of Chinese datasets in B^{2}NERD.

[a] https://www.heywhale.com/mw/dataset/617969ec768f3b0017862990/file
[b] Boson,Peopledaily2014 and Finance-sina datasets are available at https://github.com/liucongg/NLPDataSet
[c] https://www.biendata.xyz/competition/CCKS2017_2/
[d] https://www.biendata.xyz/competition/CCKS2018_1/
[e] https://www.biendata.xyz/competition/ccks_2019_1/
[f] https://www.biendata.xyz/competition/ccks_2020_2_1/
[g] https://www.heywhale.com/home/competition/620b34ed28270b0017b823ad/content/2
[h] http://www.fudan-disc.com/sharedtask/imcs21/index.html
[i] http://tcci.ccf.org.cn/conference/2018/taskdata.php
[j] https://tianchi.aliyun.com/markets/tianchi/ruijin#guid-03
[k] https://tianchi.aliyun.com/competition/entrance/531824/information
[l] https://www.heywhale.com/mw/dataset/609a27c5f29cea00179233f3/file
[m] https://www.datafountain.cn/competitions/561/datasets
[n] https://www.biendata.xyz/competition/ccks_2020_8/
[o] http://cail.cipsc.org.cn/task_summit.html?raceID=7&cail_tag=2021
[p] https://tianchi.aliyun.com/competition/entrance/531900/introduction
[q] https://blog.csdn.net/qq_36426650/article/details/87719204

person	organization	life	education
general mention of person	general mention of organization	movie actor	AI algorithm
mythical figure	organization	movie age rating	academic conference
person	organization (without political group)	movie character	academic discipline
person -> writer	organization -> university	movie director	academic journal
person -> musical artist	organization -> band	movie genre	application domain
person -> others	organization -> corporation	movie plot	scientific theory
person -> politician	organization -> group or band	movie quality rating or descriptor	experiment metrics
person -> researcher	organization -> others	movie song mention	research field
person -> scientist	organization -> political party	movie title	research task
		movie trailer or preview term	
		restaurant amenity service	
		restaurant name	
		restaurant quality descriptor	
		cuisine type	

location	object	biomedical	others
country	animal	DNA	award
exact location	astronomical object	RNA	review related term
general mention of geo-political entity	orientation of vehicle	anatomy	dish or beverage name
general mention of location -> facility	brand of vehicle	biological molecules	else
general mention of location -> others	color of vehicle	biomedical term	engineering material
geo-political entity	food items	cell line	language
geographical area	general mention of vehicle	cell type	legal document
location	general mention of weapon	chemical	literary genre type
road	machine or equipment	chemical compound	manufacturing concept or principle
location (without country)	musical instrument	chemical element	manufacturing process
location (without geo-political entity)	technological instrument	disease	manufacturing standard
location -> facility	plant	disease name	manufacturing technology
nationalities or political group	position of vehicle	enzyme name	mechanical property
proximity or location description	product name	gene	miscellaneous
river	product name -> vintage car	microorganism	music genre
	product name -> MPV	protein name	programming language
	product name -> SUV		process evaluation technique
	product name -> bus		
	product name -> coupe		
	product name -> estate car		
	product name -> hatchback		
	product name -> motorcycle		
	product name -> roadster		
	product name -> sedan		
	product name -> sports car		
	product name -> truck		
	product name -> van		
	product name -> vehicle		
	vehicle type		
	vehicle velocity		
	vehicle model		
	vehicle range		

work	event	time	metric
creative work	event name	date or period	cardinal number
creative work -> album	event name -> election	operating hours	measurement quantity
creative work -> book	event name -> geographical phenomenon	sub-day time expression	monetary amount (with unit)
creative work -> magazine	event name -> others	well-defined time interval	percentage (with %)
creative work -> media contents	event or activity name	year or time period	price description
creative work -> poem			process parameter
creative work -> song			ordinal number

Table 17: Full NER taxonomy of English entities in B^2NERD.

人物相关	地名相关	生物医学	物品相关	组织机构相关
人名	乡镇	中医证候	产品名	组织机构名
人名->体育经理	产品产地	中药功效	产品名->乐器	组织机构名(不带地理政治实体)
人名->其它	兴趣点	医学检查项目	产品名->交通工具	组织机构名->体育团队
人名->政治人物	区县	医学检查项目->实验室检验项目	产品名->其它	组织机构名->公共公司
人名->歌手	单元号	医学检查项目->影像检查	产品名->待售产品	组织机构名->公司
人名->犯罪嫌疑人	国家	医学检查项目的名称或泛称	产品名->服装	组织机构名->其它
人名->神职人员	地名	医学检测结果	产品名->相关产品	组织机构名->政府机构
人名->科学家	地名->其它	医疗设备	产品名->金融产品	组织机构名->汽车制造商
人名->艺术家	地名->景点	医院科室	产品名->食品	组织机构名->私营公司
人名->被害人	地名->目的地	微生物名	产品名->食品或饮品	组织机构名->航空航天制造商
人名->运动员	地名->设施	检查或治疗程序	产品名->饮品	组织机构名->银行
人名或昵称	地名->起点	毒品或毒品成分名	产品型号	组织机构名->音乐团体
人物	地名->车站	治疗措施(不含手术)	产品系列名	组织机构泛称
人物或团体名	地名完整描述	治疗措施(含药物)	产品配件	组织机构的名称或泛称
人群泛称	地名或地理政治实体	治疗措施->手术	品牌名	
人群类别	地点(不带地理政治实体)	治疗措施描述	商品名	
国籍	地点泛称	生化成分	手机型号	
地名->人类居住地	地点的名称或泛称	疾病名	涉案物品完整描述	
头衔	地理政治实体	疾病类别	物品的名称或泛称	
收件人或收件单位	地理政治实体泛称	疾病诊断	食品类别	
民族	城市	病因		
用户群体	子兴趣点	症状		
籍贯	开发区	症状或体征		
职业或职位	房间号	症状或体征描述		
联系人名	普通辅助定位词	细胞类型		
	村民小组	给药方式		
	村社	药品名		
	楼层号	药物		
	楼栋号	药物剂型		
	次级道路	药物名		
	次级道路门牌号	药物性味		
	省份	药物成分		
	自定义目的地	药物的名称或类别		
	距离辅助定位词	药物类别		
	路口	解剖学实体(非标准)或细胞名		
	道路	解剖部位		
	门牌号	解剖部位(含动植物)		
		身体部位		
		身体部位或身体物质		

作品相关	度量相关	教育相关	时间相关	其它
作品名	产品规格尺寸	专业名	产品使用期限	产品主题
作品名->影像作品	剂量	教育相关实体	年龄段	产品使用场所
作品名->文字作品	度量	教育背景	持续时间	产品功能描述
作品名->游戏作品	数值	数学概念	时间完整表述	产品味道
作品名->电视节目	频率	解题原理或方法	时间状语	产品图案
作品名->艺术作品	程度		时间相关信息	产品型号编号
作品名->软件	重量		时间相关表述	产品外形描述
作品名->音乐作品	财产价值(带币种)			产品服务描述
作案工具	销赃金额(带币种)			产品材料类型
副作用	现金或转账金额(带币种)			产品款式
歌曲语言				产品气味
音乐主题				产品用途
音乐列表类别				产品订阅类型
音乐风格				产品配置参数
				其它
				形容词评价
				情感类型
				抽象概念
				电脑硬件规格
				电话号码
				证书文档
				金融术语
				音乐场景
				颜色

Table 18: Full NER taxonomy of Chinese entities in B^{2}NERD.