OmniNER2025: Diverse and Comprehensive Fine-Grained NER Dataset and Benchmark for Chinese

Yong Zhou*
School of Software Technology,
Dalian University of Technology
Dalian, China
tyzy8999@gmail.com

Shuaipeng Liu*†
XiaoAI Team, Xiaomi
Beijing, China
liushuaipeng@xiaomi.com

Yunqing Li
School of Cyber Science and Technology, Beihang University
Beijing, China
sy2239117@buaa.edu.cn

Mengting Hu
College of Software, Nankai University
Tianjin, China
mthu@mail.nankai.edu.cn

Wen Dai
XiaoAI Team, Xiaomi
Beijing, China
daiwen@xiaomi.com

Xiaowei Zhao
School of Software Technology,
Dalian University of Technology
Dalian, China
xiaowei.zhao@dlut.edu.cn

Xiujuan Xu†
School of Software Technology,
Dalian University of Technology
Dalian, China
xjxu@dlut.edu.cn

Abstract

As Named Entity Recognition (NER) tasks have evolved, artificial intelligence has been widely applied in this field. However, most benchmarks are limited to English, making it challenging to replicate successful experiences in other languages. To expand NER to informal and diverse Chinese text scenarios, we have proposed a new large-scale Chinese NER dataset, **OmniNER2025**. This dataset, obtained from user posts on a popular Chinese social media platform Xiaohongshu, contains 195,568 samples and 89 categories, all manually annotated. To our knowledge, it is currently the largest Chinese open-source NER dataset in terms of sample size, category diversity, and domain coverage. This dataset is more challenging than existing Chinese NER datasets and better reflects real-world applications. The large sample size and diverse entity types provide valuable research resources. Additionally, we introduced the **ERRTA** tool for error analysis and teacher model guidance, significantly reducing model errors and improving performance. In the future, we will refine the ERRTA framework and explore optimization strategies to enhance the practical value of NER models. By releasing the OmniNER2025 dataset and introducing the ERRTA tool, we have advanced fine-grained NER research and improved model performance, promoting its application and development in real-world scenarios.

CCS Concepts

• **Computing methodologies** → **Information extraction**.

Keywords

Chinese NER Dataset, Named Entity Recognition, Error Type Analysis Tool, Teacher Model Guidance

ACM Reference Format:
Yong Zhou, Shuaipeng Liu, Yunqing Li, Mengting Hu, Wen Dai, Xiaowei Zhao, and Xiujuan Xu. 2025. OmniNER2025: Diverse and Comprehensive Fine-Grained NER Dataset and Benchmark for Chinese. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3730048

1 Introduction

Named Entity Recognition (NER) is one of the core tasks in natural language processing, aimed at identifying and classifying specific named entities in text, such as person names, locations, and organizational names[2, 22]. NER is widely applied in information retrieval, data mining, and question answering systems. Despite significant progress in this field, we have observed three main limitations in previous research for practical applications.

1. NER tasks predominantly focus on formal English texts, such as news reports, legal documents, and scientific papers. Major datasets like ACE04[4], ACE05[18], CoNLL03[14], OntoNotes 5.0[12], and GENIA[7] primarily cover formal English texts. While these datasets have significantly advanced NER technology, they are less effective in handling informal texts (e.g., social media posts, user comments, and instant messages), which often contain more fine-grained and diverse entity categories. Moreover, despite the

*Both authors contributed equally to this research.
†Corresponding author.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730048

increasing importance of Chinese NER research, the coverage of Chinese NER datasets in both formal and informal texts remains insufficient.

2. High-quality, large-scale, fully manually annotated Chinese NER datasets are still lacking. NER datasets have evolved from small-scale manual annotations[7, 14], to semi-automatic and automatic methods[4, 13], and now to using large-scale pre-trained models for annotation[6, 10, 17]. Although these models perform well in many applications and significantly save resources, they still struggle in complex NER scenarios. Additionally, high-quality, large-scale, fine-grained Chinese NER datasets are still scarce.

To address these issues, we propose OmniNER2025, a high-quality, fine-grained Chinese NER dataset. Fully manually annotated, this dataset aims to overcome the limitations of existing methods in complex scenarios, providing researchers with reliable benchmark data to advance NER tasks and advanced model research. OmniNER2025 includes 32 primary categories and 89 secondary categories, covering areas such as 3C digital products, local life, pets, travel, clothing, the internet, education, home furnishings, finance, and beauty. The dataset contains 195,568 high-quality manually annotated samples, each consisting of a sentence obtained by splitting community posts. For instance, a single post may be split into 4 to 5 sentences. Each sentence is accompanied by detailed annotation sequences, including entity text, start and end positions, and entity categories. In a given example, one category may contain multiple entities.

This dataset annotates more categories and details than other existing Chinese datasets. To our knowledge, it is the largest Chinese NER dataset in terms of sample size, category diversity, and domain coverage. Its complexity and challenge allow for better differentiation of modern model capabilities.

It is important to note that our definition of "named entity" goes beyond traditional proper nouns like person names, locations, and organizations. We also include descriptive phrases, concepts, and attributes that carry significant semantic value, especially in informal contexts such as social media posts and user comments. This broad approach allows us to capture a wide range of entities, such as product names, emotional expressions, and context-specific keywords, which are often used to describe products or experiences in user-generated content. For example, in a user review, "very tender" might be an important descriptor of food texture, just as "smoke smell" is a key attribute in evaluating indoor air quality. While these phrases are not proper nouns, they function as salient entities within their respective contexts. This is essential for capturing the nuanced, fine-grained information found in domains like consumer reviews and social interactions.

Additionally, to address the issue of suboptimal model performance in various domains, we propose ERRTA, a systematic error analysis tool designed to identify and reduce various types of errors, thereby enhancing model performance across different domains. ERRTA categorizes errors into two main types: classification errors and entity recognition errors. Classification errors occur when entity recognition is accurate, but entity classification is incorrect due to ambiguity, insufficient domain knowledge, or annotation errors. Entity recognition errors are further divided into four categories: entity boundary errors, entity omissions, extra entities, and entity splitting or merging errors.

ERRTA records the quantity of each error type and uses (Entity, Categories) pairs as the smallest analysis unit, rather than the entire sample. This approach clearly highlights the model's shortcomings in various aspects. By analyzing specific error types and receiving guidance from the teacher model, we can significantly reduce these errors, thus improving the model's overall performance. ERRTA provides a systematic framework for researchers to comprehensively assess the model's performance across different error types. Gradually reducing these errors helps enhance the model's effectiveness and reliability.

In light of these advancements, the key contributions of this study are summarized as follows:

- **High-Quality Manual Annotations:** OmniNER2025 employs fully manual annotations, ensuring each entity's annotation is rigorously reviewed and verified to avoid the noise and errors common in automated annotation methods. Additionally, to further prevent noise and errors in manual annotations, we applied cross-validation and filtering techniques. By using a small validation model to filter out simple samples, we enhanced the quality and challenge of the dataset.
- **Large-Scale, Diverse, and Broadly Covered Informal Texts:** OmniNER2025 contains 195,568 high-quality manually annotated samples, covering a large amount of informal text from social media platform Xiaohongshu. This dataset exhibits high diversity and complexity, including 32 primary entity categories and 89 secondary entity categories. Unlike traditional NER benchmarks, it extends the scope of named entity recognition to cover not only proper nouns but also descriptive phrases and attributes commonly found in user-generated content. To our knowledge, this is currently the largest Chinese open-source NER dataset in terms of sample size, category diversity, and domain coverage.
- **Systematic Error Analysis Tool ERRTA:** We introduced ERRTA (Error Type Analysis Tool), which records the number of each error type using (Entity, Categories) tuples as the smallest unit. This tool clearly showcases the model's deficiencies in various aspects. Through specific error type analysis and teacher model guidance, we can significantly reduce various types of errors and improve model performance. ERRTA provides a systematic framework, enabling researchers to comprehensively evaluate and enhance the overall performance of models.[1].

2 Related work

2.1 Development of English NER Datasets

NER research in English has extensively utilized formal texts such as news articles, legal documents, and scientific papers. The ACE04 dataset[4], part of the Automatic Content Extraction (ACE) program, includes annotations for various types of entities in formal texts, such as persons, organizations, and geopolitical entities. The CoNLL03 dataset[14] is widely recognized for its high-quality annotations of person, organization, location, and miscellaneous entities in news texts, making it a standard benchmark for English NER.

[1]Code and datasets are available athttps://github.com/TYZY89/OmniNER2025

To address limitations in domain coverage, OntoNotes 5.0 [12] introduced annotations across multiple genres, including news, telephone conversations, blogs, and broadcast transcripts, offering a broader resource for NER research. The GENIA corpus [7], a specialized biomedical dataset, provides detailed annotations for biological entities, facilitating domain-specific NER advancements. Similarly, the BC5CDR dataset [9] focuses on disease and chemical entity recognition, further pushing NER applications in biomedical research.

In recent years, social media and noisy text NER datasets have gained attention due to the growing presence of user-generated content. The WNUT16 [15] and WNUT17 [3] datasets focus on NER in noisy, informal, and user-generated texts, capturing emerging entities and dynamic language patterns.

2.2 Advances in Chinese NER Datasets

In the Chinese NER field, several important datasets have been developed, each contributing to different aspects of NER research. The MSRA dataset [8], composed primarily of news articles, remains one of the most widely used benchmarks for Chinese NER.

Beyond formal domains, the WeiboNER dataset [5, 11] was one of the first datasets to address social media NER in Chinese, capturing informal, user-generated text from Sina Weibo. For domain-specific NER, the Resume NER dataset [26] focuses on professional and recruitment documents, making it useful for applications in human resources and job-related information extraction. The CLUENER2020 dataset [22] extends general-domain Chinese NER research by introducing a broader set of entity types, which are often overlooked in traditional NER benchmarks.

2.3 Key Differences and Gaps in Existing NER Benchmarks

Key Differences and Gaps in Existing NER Benchmarks While traditional NER datasets focus on formal, well-structured text, there is a growing need for fine-grained, domain-adaptive, and context-aware NER benchmarks. Many datasets struggle to capture emerging entities, colloquial expressions, and implicit mentions—challenges that are particularly evident in user-generated content and informal discussions. Additionally, most existing NER benchmarks lack comprehensive multi-level entity annotations, limiting their effectiveness in understanding nuanced and hierarchical entity relationships.

OmniNER2025 addresses these challenges by: Extending the scope of named entity recognition to include not just proper nouns but also descriptive phrases, attributes, and contextually significant concepts in social media text. Providing high-quality, manually annotated data from Xiaohongshu, a platform rich in consumer reviews, lifestyle discussions, and informal interactions. Incorporating multi-level entity categorization, supporting fine-grained NER research across a wide range of entity types and contexts.

3 OmniNER2025 Overview

We collected community posts published by users on a social media platform Xiaohongshu that offers various channels for sharing shopping experiences, product feedback, travel guides, and lifestyle tips. Users primarily publish public posts or questions through the community feature. First, we collected these community posts and questions through platform data logs. We anonymized and sanitized private information in the raw data to ensure data security and user privacy. We used annotators to filter comments and conversations related to product issues, user experience, and security concerns, excluding parts containing taboo language and sensitive content. Finally, we retained 80,029 community posts, forming a corpus of the same size. Based on this corpus, we developed OmniNER2025, a Chinese NER dataset containing 195,568 examples, covering 32 domains including 3C digital products, local life, pets, travel, clothing, internet, education, home, finance, and beauty. These domains are divided into 89 entity types.

3.1 Data Annotation

Despite having detailed annotation guidelines, the annotation process is complex and prone to errors. To ensure accuracy and consistency, we referenced the annotation processes of ACE 2005 [4], MAVEN [20], and MUSIED [21], organizing a two-stage iterative annotation process. We recruited annotators with domain-specific knowledge and provided them with comprehensive guideline training. Annotators first underwent annotation practice, and only those with an accuracy rate above 90% proceeded to the formal annotation stage.

In the first stage, each document was annotated by three independent annotators. An annotation was considered complete only if all three annotators agreed. If two out of three annotators agreed, a voting mechanism determined the final annotation result. If all three annotators disagreed, the document was submitted to domain experts for final annotation.

In the second stage, domain experts reviewed and annotated documents with discrepancies from the first stage, ensuring the accuracy and consistency of annotations. All documents underwent this two-stage iterative annotation, resulting in a high-quality annotated dataset.

3.2 Annotation Challenges and Solutions

In NER tasks, most research focuses on named entity recognition from formal texts such as news articles, and Wikipedia documents. As user-generated texts accumulate on the internet and within enterprises, effective named entity recognition from these informal texts (often from multiple heterogeneous sources) has become increasingly important. Due to the informal nature of these texts, the annotation process needs to consider two main issues:

Boundary Confusion: The linguistic diversity and casual expressions in informal texts can easily lead to boundary confusion of entities. As shown in sentence s1 of table 1, the annotated entity is "烟味" (smoke smell) instead of "除烟味" (remove smoke smell). Although "除烟味" (remove smoke smell) might be correct in some contexts, it contains redundant information. In sentence s2, we annotate "非常嫩" (very tender) instead of "嫩" (tender), as "非常嫩" (very tender) more accurately conveys the intended meaning. For boundary confusion, our annotation principle is: avoid redundant information while fully expressing the actual situation.

Granularity Consistency: During annotation, we need to consider the level of granularity to ensure accuracy and practicality. In sentence s3, the annotation result is "改善法令纹和抬头纹" (improves nasolabial folds and forehead wrinkles). This coarse-grained

Table 1: Example Sentences with Entity Highlighting in OmniNER2025. Examples of OmniNER2025. Red and purple are used to mark entities and to distinguish linked entities.

Sentences
S1: 除烟味的同时，还能除其他异味
While removing smoke smell, it can also remove other odors
S2: 玫瑰豉油鸡 选用鸡腿肉，鸡肉非常嫩，咸鲜口，不错！
Rose soy sauce chicken uses chicken thigh meat, the meat is very tender, salty and delicious, good!
S3: 改善法令纹和抬头纹，雅萌bloom还不错
Improves nasolabial folds and forehead wrinkles, YA-MAN bloom is pretty good
S4: 撸猫上瘾 中午睡姿太迷人，忍不住又撸了
Addicted to petting the cat. Its noon nap posture was too adorable; I couldn't help but pet it again
S5: 而且没觉得有啥柔光效果呀[笑哭R]瞎买瞎用吧……怪不得有人吐槽鸡肋……滋润度是挺好
Moreover, I didn't feel any soft light effect at all [laughing and crying emoji]. Just buying and using it blindly... No wonder some people say it's useless...
The moisturizing effect is pretty good
S6: 我真的真的真的超爱各种复古风了[害羞R] 经久不衰耐看哈哈[萌萌哒R]
I really, really, really love all kinds of vintage styles [shy emoji]. They are timeless and always pleasing to the eye, haha [cute emoji]
S7: 华东理工大学教工宿舍的出租房，清爽安静，可以享受绿色的那个大操场，小区里有小河可以垂钓
The rental apartments in the faculty dormitories of East China University of Science and Technology are clean and quiet. You can enjoy
the green playground, and there is a small river in the community where you can fish
S8: 期待下次与粉丝们的见面 Nanci 的粉丝都特别好看 又有智慧 游戏环节相当有趣
Looking forward to meeting the fans next time. Nanci's fans are very attractive and intelligent. The game session was quite interesting
S9: 11月22日后MINI CLUBMAN 2.0t改款信息 #开MINI的人[话题]#
After November 22, MINI CLUBMAN 2.0t facelift information #MINI owners [topic]

annotation includes a complete effect description, clearly expressing the specific improvement goals and reflecting complete functional information. If annotated as "法令纹" (nasolabial folds) and "抬头纹" (forehead wrinkles), the overly fine-grained annotation might overlook the overall functional description, failing to fully reflect the product's efficacy. For granularity consistency, our principle is: ensure the accuracy and completeness of annotations while making them practical and comprehensible in real applications.

3.3 Data Filtering and Validation

To ensure our dataset is challenging for modern models, we applied data filtering techniques, which we call the cross-validation and filtering method. First, we split the annotated dataset into 4 folders. For each folder, we trained a small data quality validation model based on qwen2-0.5b-instruct, representing lower model capability compared to other complete models. We used the model trained from the current folder to predict the other folders and applied the same process to each folder. Finally, each folder had 3 predictions. We removed all samples correctly predicted by all 3 predictions, considering them too simple for our model. Ultimately, we filtered the annotated dataset from 195,718 samples with 90 entity categories to 195,568 samples with 89 entity categories.

3.4 Data Statistics

To better understand our dataset, we provide examples in Table 2. In the second example, the format (Entity, Category) represents the entities and their categories. For instance, the entity "三月份 (March)" corresponds to the category "时间_年/月维度时间 (Time_Year/Month Dimension)."Here, "时间 (Time)" denotes the primary category, and "年/月维度时间 (Year/Month Dimension)" denotes the secondary category. The underscore "_" separates the primary and secondary categories: before the underscore is the primary category, and after it is the secondary category. The slash "/" indicates that the entity belongs to either category. Categories without an underscore represent both primary and secondary categories. The entire dataset comprises 32 primary and 89 secondary categories.

In Table 3, we outline the statistical information of OmniNER2025 and other mainstream Chinese NER datasets. As shown, MSRANER [8] includes three basic categories: person names, locations, and organizations. Weibo NER [5, 11] adds a geopolitical category. Resume NER [26] consists of 8 categories, significantly more than MSRA, but with a highly imbalanced distribution. CLUENER2020 [22] balances and improves the data volume in each category. CMeEE focuses on biomedical scenarios. OmniNER2025 is the most comprehensive Chinese open-source NER dataset, containing 195,568 high-quality annotated samples, covering 32 primary categories and 89 secondary categories. It is the only dataset that includes both levels of categories, making it more challenging.

4 Experiments

In this study, we evaluate the OmniNER2025 on a series of publicly available Chinese pre-trained models through a process of fine-tuning. The models used include: Qwen2 Series[24]: **Qwen2-7b-instruct**, **Qwen2-1.5B-Instruct**, and **Qwen2-0.5B-Instruct**.

Table 2: Entity Category Pairs Extracted by OmniNER2025.

Sentences	Entity Category Pair
最主要的是！除烟味的同时，还能除其他异味 螺蛳粉、榴莲味、火锅烧烤味、汗味、脚臭味都可以去除。 **The main thing is! It removes smoke odor as well as other odors like snail noodles, durian, hot pot barbecue, sweat, and foot odor.**	(烟味, 气味/味道_气味), (异味, 气味/味道_气味), (螺蛳粉, 气味/味道_气味), (榴莲味, 气味/味道_气味), (火锅烧烤味, 气味/味道_气味), (汗味, 气味/味道_气味), (脚臭味, 气味/味道_气味) **(smoke odor, smell/taste_smell), (odor, smell/taste_smell),** **(snail noodles, smell/taste_smell), (durian odor, smell/taste_smell),** **(hot pot barbecue odor, smell/taste_smell), (sweat odor, smell/taste_smell),** **(foot odor, smell/taste_smell)**
三月份在金山的候就一次。以泡酒为基底，加入檬、梨、大利苦艾酒以及甜白酒。 **I ordered it once in March when I was in San Francisco. It is based on sparkling wine, with lemon, pineapple, Italian vermouth, and sweet white wine added.**	(三月份, 时间_年/月维度时间), (金山, **poi**_其他), (一次, 数量), (泡酒, 品类_产品), (檬, 品类_产品), (梨, 品类_产品), (苦艾酒, 品类_产品), (甜白酒, 品类_产品), (大利, **aoi**_产地), (加入, 行为_操作行为) **(March, Time_Year/Month Dimension), (San Francisco, poi_Other),** **(Once, Quantity), (Sparkling Wine, Category_Product),** **(Lemon, Category_Product), (Pineapple, Category_Product),** **(Vermouth, Category_Product), (Sweet White Wine, Category_Product),** **(Italy, aoi_Place of Origin), (Added, Action_Operational Action)**

Baichuan2-7B-Chat [23]. **Internlm2_5-7b-chat**[1]. BERT Series[2]: **bert-base-chinese** and **chinese-roberta-wwm-ext-large**.

4.1 Benchmark Results

We utilized the LLaMA-Factory[3] framework to train and evaluate models on the OmniNER2025 dataset. For fine-tuning on OmniNER2025, we used the following configuration: We employed the AdamW optimizer with a learning rate of 1e-4 and a cosine scheduler with 10% warmup ratio. Models were trained for 3 epochs with a per-device batch size of 2 and gradient accumulation steps of 2. Training was conducted using mixed precision (bfloat16) to improve efficiency. All models were trained using eight NVIDIA A100 80G GPUs. In terms of computational resources, full parameter fine-tuning of 7B models required approximately 65GB VRAM per GPU and took about 8 hours. Researchers with limited hardware may use gradient checkpointing to reduce memory usage at the cost of 20% longer training time.

Our experiments focused on the 15 primary categories from OmniNER2025, as previously mentioned. The results, presented in Table 4, indicate that larger pre-trained models tend to perform better. Despite the significantly larger dataset compared to other Chinese NER benchmarks, the best baseline F1 score was only 73, which is considerably lower than conventional Chinese NER tasks. For instance, MSRANER [16] achieved 95 F1, and CLUENER2020 [22] reported 80. In contrast, English NER datasets typically reach F1 scores above 85, with some exceeding 95.

4.1.1 Performance Gap Between Generative and Non-Generative Models. To better understand model performance, we also evaluated non-generative models on OmniNER2025. The F1 scores for Qwen2-7B-Instruct, Baichuan2-7B-Chat, Internlm2_5-7B-Chat, Qwen2-1.5B-Instruct, and Qwen2-0.5B-Instruct were **73.12, 72.77, 72.79, 70.84, and 67.31**, respectively. In contrast, bert-base-chinese and chinese-roberta-wwm-ext-large achieved only **51.71 and 55.18**, highlighting a significant performance gap between generative and non-generative models.

Notably, previous studies on datasets with fewer categories have shown that non-generative models can achieve competitive or even superior performance due to their efficiency in structured classification. In scenarios with a large number of categories, the importance of the rich prior knowledge embedded in generative models becomes apparent. For instance, distinguishing between highly similar categories requires strong semantic understanding, which generative models excel at due to their broader contextual awareness and knowledge learned from vast pre-training corpora.

4.2 Human Performance

To understand the complexity of Named Entity Recognition (NER) tasks and compare model performance with humans, we incorporated human evaluation into our experiments. We used a two-stage "training and evaluation" approach for this comparison. During the training phase, annotators familiarized themselves with NER categories and definitions. We selected individuals with accuracy rates above 90% for annotation. In the evaluation phase, **amateur annotators without domain knowledge labeled instances in the test set**. Similar to SuperGLUE [19] and CBLUE [25], we trained annotators before they labeled the test data. Annotators labeled development set data and validated their annotations against the gold standard. They repeatedly corrected errors to master the task. Finally, annotators labeled test data, and these annotations were used to calculate the final human score. Given the numerous NER categories and rigorous evaluation process, we averaged the scores of three annotators to calculate final human performance. In all domains, human performance surpassed that of the models, **with**

[2]https://huggingface.co/google-bert/bert-base-chinese, https://huggingface.co/hfl/chinese-roberta-wwm-ext-large
[3]https://github.com/hiyouga/LLaMA-Factory

Table 3: Attribute of OmniNER2025 Datasets.

NER Datasets (Chinese)	# Doc	# Tokens	# Sentence (Total)	# Train	# Dev	# Test	# Entity Categories
MSRA	-	2,393,207	50,729	46,364	-	4,365	3
Weibo	-	104,877	1,889	1,350	269	270	4
Resume	-	157,847	4,761	3,821	463	477	9
WNUT16	-	-	7,244	2,394	1,000	3,850	10
WNUT17	-	-	5,690	3,394	1,009	1,287	6
CLUENER2020	740,000	283,012	13,436	10,748	1,343	1,345	10
CMeEE	-	-	23,000	15,000	5,000	3,000	-
OmniNER2025 (Ours)	80,029	**5,429,969**	**195,568**	**156,454**	**19,557**	**19,557**	**89**

Table 4: We report results for 15 selected primary categories from the OmniNER2025 dataset due to space constraints. The complete experimental results, encompassing all 32 primary categories (89 secondary categories). The best F1 scores are highlighted in bold, while the worst are underlined.

Entities Categories	Qwen2-7B-Instruct			Baichuan2-7B-Chat			Internlm2_5-7B-chat			Qwen2-1.5B-Instruct			Qwen2-0.5B-Instruct		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Category	83.23	84.44	**83.83**	82.14	84.07	83.09	82.32	83.44	82.78	81.49	83.36	82.42	62.06	63.67	62.86
IP	77.81	80.42	79.11	79.17	88.37	**83.52**	77.81	80.42	79.11	70.00	75.94	72.85	50.00	60.56	54.91
Product Series	69.73	68.83	**69.27**	63.72	64.51	64.11	69.13	69.41	69.27	65.79	63.71	64.74	60.26	60.14	60.20
Ingredient	65.25	61.74	63.44	70.37	65.52	**67.86**	64.86	60.00	62.33	60.32	58.45	59.37	56.68	52.38	54.44
Material	68.24	68.56	68.40	71.60	71.60	**71.60**	71.27	70.93	71.10	67.96	66.35	67.15	62.40	60.82	61.60
Function/Effect	70.54	70.36	70.45	73.21	76.88	**75.00**	70.61	69.37	69.98	60.71	60.87	60.79	63.53	61.85	62.68
Negative Effect	32.11	24.14	27.56	33.33	40.00	**36.36**	24.77	18.62	21.26	24.59	20.00	23.87	26.61	20.00	22.83
Style/Appearance	58.20	58.98	**58.59**	56.16	51.90	53.95	56.55	56.07	56.31	55.08	50.97	52.83	55.47	50.45	52.55
Color	79.38	78.35	78.86	80.58	81.16	**80.87**	78.05	77.99	78.02	79.02	77.78	78.40	73.66	72.76	73.21
Product Specification	73.63	70.68	**72.12**	68.28	73.84	70.95	70.63	69.06	69.83	68.63	66.31	67.45	62.56	63.82	63.18
Product Feature	50.24	48.81	**49.51**	38.76	36.13	37.40	48.16	46.67	47.41	45.79	43.42	44.57	39.99	38.18	39.06
Style	64.20	63.60	63.90	68.97	71.43	**70.18**	58.20	58.98	58.59	55.62	51.45	53.47	46.10	40.88	43.33
Pattern	48.82	45.60	47.16	52.94	56.25	**54.55**	50.00	40.66	44.85	50.00	34.62	37.84	44.44	30.77	36.36
Craft	65.56	69.60	67.52	76.67	79.31	**77.97**	64.58	68.28	66.38	60.24	66.08	63.03	58.72	60.79	59.74
Smell/Taste	57.14	56.90	57.02	62.96	68.00	**65.38**	59.58	60.08	59.83	51.37	53.47	52.40	35.78	30.23	32.77
Micro Avg (89 Categories)	73.25	72.99	**73.12**	72.38	73.16	72.77	73.00	72.57	72.79	70.99	70.68	70.84	67.35	67.27	67.31

final F1 scores stabilizing around 90. The reasons for this are as follows:

1) Numerous categories: The extensive number of entity categories and their broad coverage increased the task's complexity. **2) Informal text:** The variability in wording and structure of informal text makes it more challenging for models to process, while humans perform better with informal language. **3) Conversational text:** Conversational text aligns more with human daily usage, making it more difficult for current models to process. **4) Human comprehension:** Humans have an advantage in understanding context and handling ambiguity.

Through this experiment, we observed that advanced models possess strong capabilities but also have limitations in handling complex Chinese NER tasks, particularly with numerous entity categories and conversational or informal text. We hope these findings provide valuable insights for future model improvements.

5 ERRTA (Error Type Analysis) Overview

For OmniNER2025, we conducted a case analysis by comparing the model prediction results with the manual annotations, identifying two main categories of error types: entity recognition errors and classification errors.

- **Classification Errors:** The entity recognition is correct, but due to ambiguity, lack of relevant domain knowledge, or annotation errors, the entity classification is incorrect.
- **Entity Recognition Errors:** The model's predicted entities do not completely match the manual annotations and are divided into the following four categories:
  - **Boundary Errors:** The predicted entity range does not match the actual annotation.
  - **Missing Entities:** The model fails to recognize some entities that actually exist.

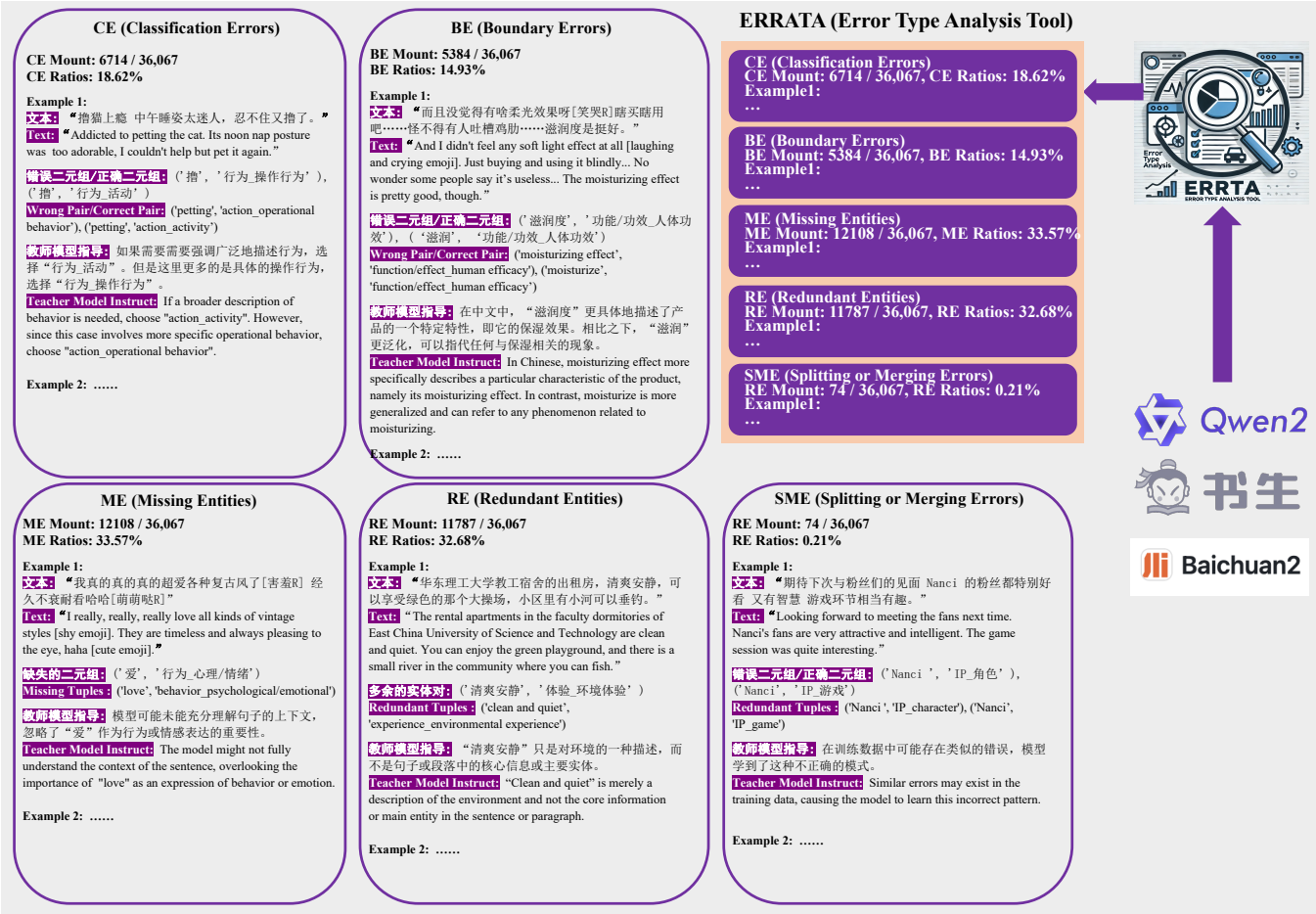

Figure 1: The overall architecture of our ERRTA.

- **Redundant Entities:** The model incorrectly recognizes entities that do not exist.
- **Splitting or Merging Errors:** The model incorrectly splits one entity into multiple entities or merges multiple entities into one.

We proposed an **ERRTA (Error Type Analysis)** to record the number of each error type, using (Entity, Categories) tuples as the smallest unit instead of the entire sample. This approach allows each sample to contain multiple errors. ERRTA documents all error types and shows the number of each type of error in the baseline model, clearly highlighting the model's deficiencies across various aspects. By leveraging specific error type analysis and guidance from the teacher model Qwen2-72B-instruct, we can significantly reduce various types of errors, thereby improving the model's performance. For instance, a decrease in "boundary errors" within a specific domain indicates the model's improved strength in that area. ERRTA is illustrated in Figure 1.

5.1 ERRTA Case Study

5.1.1 Classification Errors. As shown in sentence s4 of table 1, the prediction was "['撸', '行为_活动']([’petting’, ’action_activity’])," while the annotation was "['撸', '行为_操作行为'] ([’petting’, ’action_operational behavior’])." According to the teacher model guidance, if a broad description of behavior is needed, choose "活动 (activity)"; if more specific operational behavior is involved, choose "行为_操作行为 (operational behavior)." Here, it involves more specific operational behavior, so "行为_操作行为 (operational behavior)" should be chosen.

5.1.2 Boundary Error Case. As shown in sentence s5 of table 1, the model predicted "滋润度 (moisturizing effect)," while the manual annotation was "滋润 (moisturize)." According to the teacher model guidance and the annotation guidelines "do not include redundant information, but fully express the actual situation," in Chinese, "滋润度 (moisturizing effect)" more specifically describes a particular characteristic of the product, namely its moisturizing effect, better expressing the actual situation. ERRTA helps in unifying granularity in subsequent annotations and improving data quality.

5.1.3 Missing Entities. As shown in sentence s6 of table 1, the model's prediction missed "爱 (love)" compared to the annotation. Through teacher model guidance, the baseline model might not have fully understood the context of the sentence, overlooking "爱

ERRTA-Guided Prompt Optimization

```
ROLE DEFINITION:
You are an advanced AI model specifically designed for precise named entity recognition and classification in text. Named entities include
{All Categories}. Follow the steps below to ensure accurate entity extraction and categorization.

Step 1: Understanding Common Error Types
Before making predictions, the model must first recognize five common error types and apply teacher model corrections to avoid them.

1. Classification Error (CE): The entity is recognized correctly, but its classification is incorrect.
      ● Example: The model labels ["Chanel", "Product Category"], but the correct classification is ["Chanel", "Brand"].
      ● Correction: "Chanel" refers to a brand, not a product category.

2. Boundary Error (BE): The extracted entity span does not match the full correct phrase.
      ● Example: The model extracts ["Shampoo", "Product Category"], but the correct entity is ["Gentle Shampoo", "Product Category"].
      ● Correction: Ensure the full entity phrase is extracted without truncation.

3. Missing Entity (ME): The model fails to recognize an entity that should be present.
      ● Example: The model does not recognize ["Hairball Syndrome", "Pet Health Issue"].
      ● Correction: Improve recall by ensuring all meaningful entities in the text are identified.

4. Redundant Entity (RE): The model incorrectly recognizes non-entity words as named entities.
      ● Example: The model extracts ["Beautiful", "Style"], but this is not a named entity.
      ● Correction: Filter out adjectives, verbs, or general descriptive words.

5. Splitting/Merging Error (SME): The model incorrectly splits or merges entities.
      ● Example: The model outputs [["2024", "Time"], ["Olympics", "Event"]], but the correct entity should be [["2024 Olympics", "Event"]].
      ● Correction: Ensure multi-word entities are extracted as a whole without improper splitting.

Step 2: Error Self-Check Mechanism
Before generating the final output, the model must perform the following validation steps:
● Is the entity span complete? If not, ensure full phrase extraction.
● Is the entity classification accurate? If multiple categories apply, choose the most specific and correct one.
● Are all important entities included? Ensure no key entities are omitted.
● Are redundant entities removed? Filter out adjectives, verbs, or irrelevant words.
● Are entities properly split or merged? Ensure multi-word entities are extracted as a whole.

Identify all named entities in the text. Classify each entity into one of the predefined categories. If there are no identifiable named
entities in the text, return an empty list. Input format: text data, which may come from different fields and backgrounds. Output format: a
list containing named entities and their categories. Below is the input data, please give the output result:

{input_text}
```

Figure 2: ERRTA-Guided Prompt Optimization.

(love)" as an important expression of behavior or emotion. Through missing entity case analysis and statistical analysis, we can pass the ERRTA missing entity analysis results to manual re-evaluation to determine whether "爱 (love)" needs to be annotated.

5.1.4 Redundant Entities. As shown in sentence s7 of table 1, the model predicted "清爽安静 (clean and quiet)," while the manual annotation did not include this entity. According to the teacher model guidance, "清爽安静 (clean and quiet)" is merely a description of the environment and not the core information or main entity in the sentence or paragraph. Through ERRTA's redundant entity analysis, we can better understand which descriptions are secondary and further improve the model's performance in different contexts, thereby enhancing overall data processing and model performance.

5.1.5 Splitting or Merging Errors. As shown in sentence s8 and s9 of table 1, the model predicted "Nanci " and "2.0t" respectively, while the actual results were "Nanci" and "2.0t." Due to the structured format of the results, the former and latter had extra spaces at the beginning and end, respectively, and were judged as errors. Through splitting or merging error analysis, we can improve such errors, such as removing leading and trailing spaces in instructions or using unstructured results. Another situation is that similar errors might exist in the training data, and the model has learned this incorrect pattern. Therefore, according to ERRTA, we can prevent the occurrence of splitting or merging errors.

5.2 ERRTA for Model Improvement

The ERRTA framework is designed to systematically analyze and improve NER model performance by identifying and reducing different types of errors. It consists of several key components, as shown in Figure 1:

1) Error Type Statistics: By counting each error type, we can compare model performance across different categories and prioritize improvements.

2) Error Example Display: Listing typical error cases provides insights into error patterns, facilitating targeted troubleshooting.

3) Error Distribution Analysis (Teacher Model Guidance): Analyze the distribution of different types of errors in various entity categories to help researchers understand which entity categories the model performs poorly on and then optimize the recognition and classification of these categories. This analysis, guided

by teacher models, enables a deeper understanding ofmodel performance across different contexts and domains.

5.2.1 ERRTA-Guided Prompt Optimization. ERRTA identifies five common error types in Named Entity Recognition (NER): **Classification Errors (CE), Boundary Errors (BE), Missing Entities (ME), Redundant Entities (RE), and Splitting/Merging Errors (SME)**. To systematically reduce these errors, we introduce the following structured prompt optimization framework.

Role Definition. You are an advanced AI model specifically designed for precise named entity recognition and classification in text. Named entities include {All Categories}. Follow the steps below to ensure accurate entity extraction and categorization.

Step 1: Understanding Common Error Types. To improve entity recognition accuracy, we categorize errors into five common types: Classification Errors (CE), Boundary Errors (BE), Missing Entities (ME), Redundant Entities (RE), and Splitting/Merging Errors (SME). These error types are thoroughly analyzed in ERRTA Case Study, where concrete examples illustrate their impact.

Our prompt optimization approach is designed to mitigate these errors by incorporating structured guidance based on error characteristics. Instead of redundantly listing examples, we systematically integrate key error prevention strategies into the prompt. For a detailed analysis of how these errors manifest and the corrections provided by the teacher model, refer to **ERRTA Case Study**.

Step 2: Error Self-Check Mechanism. Before generating the final output, the model must perform the following validation steps:

- **Is the entity span complete?** If not, ensure full phrase extraction.
- **Is the entity classification accurate?** If multiple categories apply, choose the most specific and correct one.
- **Are all important entities included?** Ensure no key entities are omitted.
- **Are redundant entities removed?** Filter out adjectives, verbs, or irrelevant words.
- **Are entities properly split or merged?** Ensure multi-word entities are extracted as a whole.

For the complete structured prompt, please refer to Figure 2.

5.2.2 Reduction in Error Counts. To evaluate the impact of ERRTA-driven improvements, we conducted an error analysis before and after applying ERRTA-based refinements using Qwen2-7B-Instruct as a case study. Table 5 summarizes the changes in error counts across different error types. The reduction in error rates across different error categories indicates that ERRTA-guided prompt engineering effectively reduces common sources of NER errors in this specific model.

5.2.3 Improvement in Model Performance. We further evaluated the impact of ERRTA on model performance by comparing F1 scores before and after applying ERRTA-driven refinements. The results are presented in Table 6.

After integrating ERRTA-based optimizations, all five models demonstrated consistent improvements in F1-score, with gains ranging from +2.11 (Qwen2-0.5B-Instruct) to +2.72 (Qwen2-7B-Instruct and InternLM2_5-7B-Chat). The integration of ERRTA into prompt engineering presents a lightweight yet highly effective method for improving generative NER models without requiring additional training data.

Table 5: Reduction in error counts across different error types after applying ERRTA. The total number of errors reduced is approximately 3,650, based on the Qwen2-7B-Instruct model.

Error Type	Before ERRTA	After ERRTA
Classification Error	6,714	5,745 (-14.4%)
Boundary Error	5,384	4,730 (-12.2%)
Missing Entity	12,108	10,870 (-10.2%)
Redundant Entity	11,787	10,986 (-6.8%)
Splitting/Merging Error	74	60 (-18.9%)

Table 6: NER F1-score comparison on the OmniNER2025 dataset before and after applying ERRTA.

Model	Before ERRTA	After ERRTA
Qwen2-7B-Instruct	73.12	75.84 (+2.72)
Baichuan2-7B-Chat	72.77	75.46 (+2.69)
InternLM2_5-7B-Chat	72.79	75.51 (+2.72)
Qwen2-1.5B-Instruct	70.84	73.30 (+2.46)
Qwen2-0.5B-Instruct	67.31	69.42 (+2.11)

6 Conclusion

In this work, we released the OmniNER2025 dataset, which includes more entity categories and broader coverage than existing Chinese NER datasets, based on real industry data. Our experiments with a state-of-the-art generative baseline highlight the challenges and improvement opportunities in fine-grained NER. Additionally, we introduced ERRTA for systematic error type analysis and teacher model guidance. ERRTA helps identify and reduce errors, significantly enhancing model performance. Future work will focus on refining the ERRTA framework to further improve NER models' practical application value. By providing the OmniNER2025 dataset and ERRTA tool, we advance fine-grained NER research and promote model performance through detailed error analysis and optimization, supporting real-world applications.

Acknowledgments

This work is funded in part by the National Natural Science Foundation of China Project (No. 62372078). Separately, we thank our industrial partners for providing the real-world data used in the OmniNER2025 dataset. We also appreciate the valuable feedback from anonymous reviewers and the support from our institutions.

References

[1] 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, and et al. 2024. InternLM2 Technical Report. *CoRR* abs/2403.17297 (2024). doi:10.48550/ARXIV.2403.17297 arXiv:2403.17297

[2] Kedi Chen, Jie Zhou, Qin Chen, Shunyu Liu, and Liang He. 2024. A Regularization-based Transfer Learning Method for Information Extraction via Instructed Graph Decoder. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, 1472–1485. https://aclanthology.org/2024.lrec-main.131

[3] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (Eds.). Association for Computational Linguistics, 140–147. doi:10.18653/V1/W17-4418

[4] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2004/summaries/5.htm

[5] Hangfeng He and Xu Sun. 2017. F-Score Driven Max Margin Neural Network for Named Entity Recognition in Chinese Social Media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, 713–718. doi:10.18653/V1/E17-2113

[6] Hyunjae Kim, Jaehyo Yoo, Seunghyun Yoon, and Jaewoo Kang. 2023. Automatic Creation of Named Entity Recognition Datasets by Querying Phrase Representations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 7148–7163. doi:10.18653/V1/2023.ACL-LONG.394

[7] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology, June 29 - July 3, 2003, Brisbane, Australia*. 180–182. http://bioinformatics.oupjournals.org/cgi/content/abstract/19/suppl_1/i180?etoc

[8] Gina-Anne Levow. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. In *Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia, July 22-23, 2006*, Hwee Tou Ng and Olivia O. Y. Kwong (Eds.). Association for Computational Linguistics, 108–117. https://aclanthology.org/W06-0115/

[9] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation* 2016 (2016). doi:10.1093/DATABASE/BAW068

[10] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). doi:10.48550/ARXIV.2303.08774 arXiv:2303.08774

[11] 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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 548–554. doi:10.18653/V1/D15-1064

[12] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, Julia Hockenmaier and Sebastian Riedel (Eds.). ACL, 143–152. https://aclanthology.org/W13-3516/

[13] Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively Multilingual Transfer for NER. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 151–164. doi:10.18653/V1/P19-1015

[14] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, Walter Daelemans and Miles Osborne (Eds.). ACL, 142–147. https://aclanthology.org/W03-0419/

[15] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 Named Entity Recognition Shared Task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text, NUT@COLING 2016, Osaka, Japan, December 11, 2016*, Bo Han, Alan Ritter, Leon Derczynski, Wei Xu, and Tim Baldwin (Eds.). The COLING 2016 Organizing Committee, 138–144. https://aclanthology.org/W16-3919/

[16] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 8968–8975. doi:10.1609/AAAI.V34I05.6428

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023). doi:10.48550/ARXIV.2302.13971 arXiv:2302.13971

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

[19] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 3261–3275. https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

[20] Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1652–1671. doi:10.18653/V1/2020.EMNLP-MAIN.129

[21] Xiangyu Xi, Jianwei Lv, Shuaipeng Liu, Wei Ye, Fan Yang, and Guanglu Wan. 2022. MUSIED: A Benchmark for Event Detection from Multi-Source Heterogeneous Informal Texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 2947–2964. doi:10.18653/V1/2022.EMNLP-MAIN.191

[22] Liang Xu, Yu Tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. CLUENER2020: Fine-grained Named Entity Recognition Dataset and Benchmark for Chinese. *CoRR* abs/2001.04351 (2020). arXiv:2001.04351 https://arxiv.org/abs/2001.04351

[23] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open Large-scale Language Models. *CoRR* abs/2309.10305 (2023). doi:10.48550/ARXIV.2309.10305 arXiv:2309.10305

[24] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).

[25] 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. 2022. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 7888–7915. doi:10.18653/V1/2022.ACL-LONG.544

[26] Yue Zhang and Jie Yang. 2018. Chinese NER Using Lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 1554–1564. doi:10.18653/V1/P18-1144