Overview of EvaHan2025: The First International Evaluation on Ancient Chinese Named Entity Recognition

Bin Li[1,2]✉, Bolin Chang[1,2], Ruilin Liu[3], Xue Zhao[3], Si Shen[4], Lihong Liu[5], Yan Zhu[5], Zhixing Xu[1,2], Weiguang Qu[6,2], Dongbo Wang[3,2]

[1]School of Chinese Language and Literature, Nanjing Normal University, China,
[2]Center for Language Big Data and Computational Humanities, Nanjing Normal University, China,
[3]College of Information Management, Nanjing Agricultural University, China,
[4]School of Economics and Management, Nanjing University of Science and Technology, China,
[5]Institute of Information on Traditional Chinese Medicine,
China Academy of Chinese Medical Science, China,
[6]School of Computer and Electronic Information, Nanjing Normal University, China
Correspondence: libin.njnu@gmail.com

Abstract

Ancient Chinese books have great values in history and cultural studies. Named entities like person, location, time are crucial elements, thus automatic Named Entity Recognition (NER) is considered a basic task in ancient Chinese text processing. This paper introduces EvaHan2025, the first international ancient Chinese Named Entity Recognition bake-off. The evaluation introduces a rigorous benchmark for assessing NER performance across historical and medical texts, covering 12 named entity types. A total of 13 teams participated in the competition, submitting 77 system runs. In the closed modality, where participants were restricted to using only the training data, the highest F1 scores were 85.04% on *TestA* and 90.28% on *TestB*, both derived from historical texts, compared to 84.49% on medical texts (*TestC*). The results indicate that text genre significantly impacts model performance, with historical texts generally yielding higher scores. Additionally, the intrinsic characteristics of named entities also influence recognition performance. It remains challenging to further enhance model recognition performance and to effectively integrate entities from different annotation schemes into a unified system.

1 Introduction

The EvaHan series represents an international endeavor focusing on the advancement of information processing for ancient Chinese texts. In 2022, EvaHan was convened in Marseille, France, where it conducted evaluations on word segmentation and part-of-speech tagging in ancient Chinese, contributing to the field's fundamental tasks (Li et al., 2022). The following year, the series moved to Macao, China, extending its scope to include evaluations on ancient Chinese machine translation, a significant step in computational linguistics for historical languages (Wang et al., 2023). The following year 2024, the series moved to Turin, Italy, extending its scope to include evaluations on ancient Chinese sentence segmentation and punctuation, aiming to address a critical and yet under-explored area in the processing of classical texts (Li et al., 2024). In 2025, EvaHan is set to pioneer a new frontier with its first campaign specifically devoted to the evaluation of ancient Chinese named entity recognition, aiming to enhance the identification and categorization of proper names, places, and temporal expressions in historical and medical texts, thereby fostering deeper insights into ancient Chinese text analysis.

Named Entity Recognition (NER) is a fundamental task in natural language processing that involves identifying and classifying entities (Rau, 1991). NER plays a crucial role in ancient Chinese natural language processing (NLP), facilitating the structuring and analysis of historical texts (Zhang and Yang, 2018; Li Dongmei et al., 2022). Consequently, accurate named entity recognition is essential for various downstream applications, including historical knowledge extraction, document retrieval, and the construction of large-scale historical knowledge graphs (Goyal et al., 2018; Liu Liu and Wang Dongbo, 2018). However, unlike English, ancient Chinese texts lack ex-

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 97–105

plicit word boundaries. Different from modern Chinese, ancient Chinese texts use traditional characters with a significantly larger set of characters. Additionally, the vocabulary and grammar of ancient Chinese differ from those of modern Chinese, further complicating tasks such as Named Entity Recognition (NER) and making it a particularly challenging endeavor.

The existing studies on ancient Chinese NER face several issues and challenges. First, the ancient Chinese NER mainly focused on historical texts, other types of texts are not well considered. Second, different corpora have different types of named entities. For example, historical texts include persons, locations and temporal expressions, while the medical texts have more entities like illness, cures, and formula. Third, annotation guidelines and tag set are different caused by different system developers. There is not a full named entity hierarchy for ancient Chinese. Each corpus only focus on its own interest. Thus, it is difficult to construct a wide-coverage NER system. Forth, the evaluation of ancient Chinese NER is not well set yet. The basic unit for calculation of Precision and Recall rate to be a character or an entity is still a problem, thus making it hard to compare the performances of different NER systems.

EvaHan2025 is designed as a comprehensive evaluation benchmark to address these issues. The evaluation aims to answer four key questions:

(1) How do different types of ancient Chinese texts influence NER performance?

(2) Is it possible to build an integrated system capable of handling multiple text types and multiple entity categories?

(3) Can large language models effectively generalize across different classical Chinese domains?

(4) How can we ensure a fair and unbiased evaluation, given that many pretraining corpora contain historical texts?

EvaHan2025 collects a dataset of 12 types of named entities from history and medical texts, which is designed to test the NER systems' performance on different genres and entities. And the basic unit for evaluation is the whole named entity, not the character. Considering the fast development of large language models (LLMs) , we encourage the participants to use LLMs as well as traditional models.

Entity	Meaning	Example	Dataset
NR	Person Name	蘇秦	A B
NS	Geographical Location	長平	A B
NB	Book Title	易	A
NO	Official Title	中大夫	A
NG	Country Name	秦	A
T	Time Expression	三十四年	A B
ZD	Traditional Chinese Medicine Disease	金疮	C
ZZ	Syndrome	脾胃虚弱	C
ZF	Chinese Medicinal Formula	当归散	C
ZP	Decoction Pieces	当归	C
ZS	Symptom	烦满	C
ZA	Acupoint	承扶	C

Table 1: 6 Entities involved in the evaluation

EvaHan2025 is proposed as part of the The Second Workshop on Ancient Languages Processing, co-located with The 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics. The benchmark, scoring methodology, and detailed annotation guidelines are publicly available in our GitHub repository[1], providing an open and transparent evaluation framework for the research community.

2 Task

In the EvaHan2025 evaluation task, participants are required to develop systems that automatically identify and label named entities within ancient Chinese texts, transforming raw unstructured text into structured data with entity annotations.

The evaluation focuses on 12 distinct types of named entities, covering key categories relevant to both historical texts and traditional Chinese medicine texts. Table 1 lists 12 entity types, including Person Name (NR), Geographical Location (NS), etc. Systems are assessed based on their ability to accurately detect entity boundaries and correctly classify entity types.

3 Dataset

Ancient Chinese texts, covering both historical records and Traditional Chinese Medicine literature. All the data has been annotated and proofread by experts of ancient Chinese language.

3.1 Data Source

The EvaHan2025 dataset is designed to evaluate NER performance in ancient Chinese

[1]https://github.com/GoThereGit/EvaHan

Datasets	Genre	#Char Tokens	#Entity Tokens
A	History	178167	19070
B	History	115090	11931
C	Medicine	151703	11967

Table 2: Size of each dataset

texts, covering both historical records and Traditional Chinese Medicine (TCM) literature. The dataset consists of three subsets (A, B, C), each sourced from distinct domains.

DatasetA is made of historical texts extracted from *Shiji*(史記)[2], with 6 types of named entities, developed by Nanjing Normal University.

DatasetB is also historical text extracted from *Twenty-Four Histories*(二十四史) [3], with 3 types of named entities, developed by Nanjing Agriculture University.

DatasetC is extracted from classical Traditional Chinese Medicine (TCM) texts, including TCM ancitent books such as *Liu Juanzi Guiyi Fang* (劉涓子鬼遺方)[4]. It has 6 types of entities, annotated by institute of information on traditional Chinese medicine.

Table 2 presents the size of each dataset, where Dataset A is the largest, while Dataset B is the smallest.

3.2 Data Format

All datasets are provided in plain text format, encoded in UTF-8, and include characters, punctuation marks, and a dual-layer entity annotation scheme. This dual-layer annotation structure encodes two crucial types of information: position information to indicate a character' s placement within an entity and entity type to specify its semantic category. To represent position information, the dataset employs the BMES (Beginning, Middle, End, Single) tagging scheme, which is widely used for sequence labeling tasks. In this scheme, the B (Beginning) tag marks the first character of a multi-character entity, the M (Middle) tag is assigned to characters occurring within the entity, the E (End) tag denotes the final character, and the S (Single) tag is used for entities that consist of only a single character.

By utilizing this structured annotation method, the dataset provides a clear and systematic framework for entity recognition, allowing models to effectively learn both entity boundaries and entity types.

3.3 Training Data

The training set comprises 80% of the total dataset, ensuring sufficient data for model learning.

3.4 Test Data

The test data, comprising 20% of each dataset, serves as a benchmark for evaluating system performance in NER on ancient Chinese texts. Like the training data, the test sets contain annotated entities, but they were not accessible to participants during model training, ensuring an unbiased evaluation.

Given that Datasets A and B belong to the historical text category, they provide a strong basis for assessing system performance on historical texts. In contrast, Dataset C, sourced from Traditional Chinese Medicine texts, allows for a dedicated evaluation of NER models in medical literature, which poses distinct challenges due to its specialized terminology and unique linguistic structures.

Historical texts are commonly used in ancient Chinese NER tasks and constitute a major portion of the pretraining corpora for ancient Chinese large language models. As a result, entity recognition in historical texts is typically less challenging, and models tend to achieve higher accuracy on such data.

To rigorously assess NER capabilities in historical texts, Dataset A and Dataset B are deliberately distinguished despite both belonging to the same genre. Dataset B, sourced from *The Twenty-Four Histories*, includes only three entity types, offering a comparatively simpler entity distribution. In contrast, Dataset A, contains six types of named entities, making it richer and more complex in annotation. This differentiation increases annotation complexity and introduces a higher degree of difficulty in recognizing named entities, thereby enhancing the evaluation depth of the benchmark. This distinction ensures a more precise measurement of model performance and highlights potential areas for improvement in the recognition of historical named entities.

[2]Also known as *Records of the Grand Historian*, https://en.wikipedia.org/wiki/Shiji

[3]https://en.wikipedia.org/wiki/Twenty-Four_Histories

[4]https://en.wikipedia.org/wiki/Liu_Juanzi_Guiyi_Fang

Limits	Closed Modality	Open Modality
Machine learning algorithm	No limit	No limit
Pretrained model	Only *GujiRoBERTa_jian_fan*	No limit
Training data	Only Train	No limit
Features used	Only from Train	No limit
Manual correction	Not allowed	Not allowed

Table 3: Limitations on the two modalities

4 Evaluation

Initially, each team could only access the training data. Later, the unlabeled test data was released. After the submission, the labels for the test data were also released.

4.1 Scoring

The scorer employed for EvaHan is a modified version of the one developed from SIGHAN2008 (Jin and Chen, 2008). The evaluation aligned the system-produced sentences to the gold standard ones. Then, the performance of NER were evaluated by precision, recall and F1 score. In the scoring process, we assess the correctness of entities directly, rather than Chinese characters as done in previous researches. The final ranking was based on F1 score of NER.

4.2 Two Modalities

Each participant can submit runs following two modalities. In the closed modality, the resources each team could use are limited. Each team can only use the Training data, and *GujiRoBERTa_jian_fan*[5], a large language model pretrained on a very large corpus of traditional Chinese collection, including *Siku Quanshu* (四庫全書)[6] and *Daizhige* (殆知閣)[7]. Other resources are not allowed in the closed modality.

In the open modality, there is no limit on the resources, data and models. Annotated external data, such as the components or Pinyin of the Chinese characters, word embeddings can be employed, as shown in Table 3. But each team has to state all the resources, data and models they use in each system in the final report.

[5]https://huggingface.co/hsc748NLP/GujiRoBERTa_jian_fan

[6]https://en.wikipedia.org/wiki/Siku_Quanshu

[7]https://github.com/up2hub/daizhige

4.3 Procedure

Training data was released for download from January 15, 2025. Test data was released on February 15, 2025, and results were due on 00:00 (UTC) February 21, 2025.

5 Participants and Results

5.1 Participants

A total of 23 teams registered for the task, and 13 of them submitted 77 running results. Table 4 presents the details of the participating teams. Submissions were primarily concentrated in the closed modality, while there were relatively fewer submissions in the open modality. It is important to mention that lots of submissions were initially presented in incorrect formats. It is caused by the over-generation of large language models. These errors were subsequently rectified automatically to facilitate accurate evaluation.

5.2 Results

Tables from 5 and 8 list the performance of the participating teams, arranged in descending order of the F1 scores. The Precision, Recall and F1 score for Named Entity Recognition are abbreviated as P, R and F. We classified the submissions into four categories: *TestA* and *TestB* Closed, *TestA* and *TestB* Open, *TestC* Closed, and *TestC* Open. This distinction was made because *TestA* and *TestB* consist of historical texts, whereas *TestC* is derived from Traditional Chinese Medicine texts, allowing for a comparative evaluation of NER performance across different domains. Most teams participated in the closed tests.

The highest F1 scores on *TestA* and *TestB* are 85.04% and 90.28% in the closed modality. In the open modality, they are 84.11% and 89.64%.

Since *TestA* contains a greater variety of entity categories compared to *TestB*, the performance on *TestA* is generally lower than on *TestB*. For instance, NJU achieved 88.97% and 89.64% in the closed and open modalities, respectively, on *TestB*. However, on *TestA*, NJU' s scores dropped to 83.02% and 84.11% in the closed and open modalities, respectively, reflecting a nearly 5 points decrease compared to *TestB*.

ID	Name	Affiliation	Close	Open
1	BUPT	Beijing University of Posts and Telecommunications	5	0
2	ECNU	East China Normal University	0	2
3	EPHE	École pratique des hautes études	0	1
4	HUST	Huazhong University of Science and Technology	0	3
5	NFU1	Northeast Forestry University	1	0
6	NFU2	Northeast Forestry University	3	0
7	NJU	Nanjing University	0	0
8	RUC	Renmin University of China, Midu Technology Co., Ltd.	15	18
9	SXU	Shanxi University	4	0
10	TJU	Tongji University	4	0
11	UM	University of Macau	5	0
12	UT	University of Toronto	5	1
13	WHU	Wuhan University	4	0

Table 4: Participating teams by modality

Team	TestA			TestB		
	P	R	F	P	R	F
RUC	88.97	81.45	85.04	90.22	90.34	90.28
WHU	87.23	80.65	83.81	89.47	89.92	89.70
NJU	86.64	79.69	83.02	88.73	89.21	88.97
SXU	86.30	78.78	82.37	87.43	90.09	88.74
UT	86.42	76.54	81.18	89.80	87.59	88.68
NFU1	90.77	76.75	83.17	88.42	88.75	88.59
BUPT	88.16	76.38	81.84	86.87	90.09	88.45
NFU2	89.13	79.32	83.94	89.34	87.30	88.31
UM	84.42	73.86	78.79	86.65	85.71	86.18
TJU	65.89	70.92	68.31	70.11	71.14	70.62

Table 5: Results on TestA and TestB in closed modality (%)

Team	TestA			TestB		
	P	R	F	P	R	F
NJU	88.07	80.49	84.11	90.11	89.17	89.64
UT	86.12	76.91	81.25	86.28	89.05	87.64
ECNU	83.46	75.52	79.29	89.41	85.09	87.20
HUST	83.68	73.70	78.37	88.44	84.09	86.21
RUC	73.14	84.13	78.25	82.41	82.17	82.29
EPHE	82.16	78.51	80.30	61.29	71.80	66.13

Table 6: Results on TestA and TestB in open modality (%)

Team	P	R	F
RUC	81.33	87.91	84.49
UT	82.26	84.32	83.28
NFU2	78.37	86.32	82.15
NJU	77.63	86.14	81.66
WHU	76.52	86.82	81.35
NFU1	75.58	87.36	81.05
SXU	75.91	86.09	80.68
BUPT	75.57	85.50	80.23
UM	70.33	83.09	76.18
TJU	44.04	56.77	49.60

Table 7: Results on TestC in closed modality (%)

Team	P	R	F
NJU	78.33	86.77	82.34
RUC	73.99	88.82	80.73
UT	75.35	84.05	79.46
HUST	71.32	84.32	77.28
ECNU	82.19	69.23	75.15
EPHE	46.85	59.18	52.30

Table 8: Results on TestC in open modality (%)

For *TestC*, which derived from the less common domain of Traditional Chinese Medicine texts, the scores were approximately 6 points lower than those on *TestB*. The highest F1 score of *TestC* is 84.49% in the closed modality. In the open modality, it is 82.34%.

5.3 Baselines

To provide a basis for comparison, we computed the baseline scores for each of the test sets. The baseline for ancient Chinese

Test Set	P	R	F
TestA	85.90	77.50	81.48
TestB	87.09	87.92	87.50
TestC	71.84	72.95	72.40

Table 9: Baselines (%)

Test Set	P	R	F
TestA	88.97(+3.07)	81.45(+3.96)	85.04(+3.56)
TestB	90.22(+3.14)	90.34(+2.42)	90.28(+2.78)
TestC	81.33(+9.48)	87.91(+14.95)	84.49(+12.1)

Table 10: The improvement of the best system with respect to the baseline (%)

named entity recognition was constructed using SikuRoBERTa-BiLSTM-CRF model, as shown in Table 9.

The scores of most teams exceed the baselines. The best scores from RUC outperform the baselines by around 10 points as shown in Table 10.

6 Error Analysis and Discussion

By analyzing the errors in the participating teams' systems, we can further discuss aspects related to the dataset, entity types, and large language models.

6.1 Unbalanced training samples

Based on the scores of each team across the three test sets, as shown in Tables 11 to 12, it is evident that most teams performed best on *TestB*, followed by *TestA*, while performance on *TestC* was significantly lower. This trend can be attributed to two key factors.

Firstly, while both *TestA* and *TestB* belong to the historical text category and have similar training set sizes, they differ in entity complexity. *TestB* contains only three entity types: *NR*, *NS*, and *T*, whereas *TestA* includes these three as well as three additional categories: *NO*, *NB*, and *NG*. The inclusion of these extra entity types increases the difficulty of entity recognition.

Secondly, unlike *TestA* and *TestB*, which originate from historical texts, *TestC* is sourced from Traditional Chinese Medicine literature, presenting a distinct linguistic challenge. The GujiRoBERTa model used in this evaluation was pretrained primarily on historical texts, as such texts are more commonly available. In contrast, TCM texts are relatively rare in pretraining corpora, resulting in weaker model performance on entity recognition in TCM texts compared to historical texts. This finding underscores the critical role of pretraining in large language models —a broader and more diverse pretraining corpus can significantly improve model robustness across different text domains in downstream tasks. Expanding the variety of pretraining data could enhance the model' s ability to adapt to diverse text types, leading to more consistent performance across different genres.

6.2 Entities of different datasets

Table 11 lists the quantity of annotations and corresponding scores for different entities predicted by the highest-scoring system in close modality submissions by RUC. Table 11 presents the evaluation results obtained by merging *TestA*, *TestB*, and *TestC* into a combined test set, *TestTotal*. In Table 11, *TrainTotal (Total)* means the number of gold entities in train data, *TestTotal (Gold)* means the number of gold entities in *Test Total*. *Machine (Total)* means the total number of entities tagged by the RUC' s system running on Test sets. *Machine (Correct)* means the number of correct entities tagged by RUC' s system. It is evident that *T* exhibits the highest performance, while *NB* less satisfactorily. There are two main issues with the system' s performance in NER.

Firstly, the system's performance in entity recognition is closely correlated with the frequency of entities in the training data. According to Table 11, entities with higher scores, such as *NR* and *ZP*, which achieved 90.67% and 90.24%, respectively, also appear more frequently in the training set, with occurrences of 12,968 and 4,983, respectively. Conversely, entities that are less frequent in the training data tend to have lower recognition accuracy. For example, the *NB* entity appears only 61 times in the training set, making it significantly underrepresented. As a result, the model struggles to effectively learn its patterns, leading to a much lower performance, with a score of only 50%.

Secondly, the system' s entity recognition performance is also influenced by the intrinsic characteristics of the entities themselves,

Entity	P (%)	R (%)	F (%)	Train (Total)	Test (Gold)	Machine (Correct)	Machine (Total)
T	91.37	90.68	91.02	3,452	1,062	963	1,054
NR	92.87	88.58	90.67	12,968	1,734	1,536	1,654
ZP	86.20	94.68	90.24	4,983	1,128	1,068	1,239
ZF	86.85	90.08	88.44	1,073	242	218	251
ZA	89.55	85.38	87.41	1,111	301	257	287
NS	84.49	83.36	83.92	5,550	1,124	937	1,109
NO	90.14	77.11	83.12	1,318	249	192	213
ZZ	64.77	82.61	72.61	536	69	57	88
ZD	63.69	79.72	70.81	640	143	114	179
NG	74.12	64.95	69.23	3,380	97	63	85
ZS	65.87	69.40	67.59	1,424	317	220	334
NB	100.00	33.33	50.00	61	6	2	2

Table 11: NER scores by RUC

which can either simplify or complicate the learning process. Even if an entity type is not highly frequent in the training set, its score may still be relatively high if it exhibits consistent and structured patterns. For instance, the *T* entity appears only 3,452 times in the training set but achieves a remarkably high score of 91.02%, the highest among all entity types. This is because *T* entities, unlike other entity categories, typically follow limited and highly regular forms, making them easier for models to learn. Examples include 今 (today), 冬 (winter), and 十年 (a decade).

Additionally, the *ZA* entity, despite being relatively infrequent in the training data, also achieves a high recognition score. This can be attributed to the fact that many instances of *ZA* entities appear in continuous sequences within the training data, and these sequences tend to have fixed-length structures, making them easier for models to identify. For example, in the phrase:

”循 [商阳/ZA][二间/ZA][三间/ZA] 而行，历 [合谷/ZA][阳溪/ZA] 之俞，过 [偏历/ZA][温溜/ZA] 之滨，[下廉/ZA][上廉/ZA]”

the *ZA* entities appear in a structured, repetitive format, allowing the model to recognition them with greater ease, leading to higher accuracy scores.

6.3 Character Discrepancies Due to Large Language Models

Large language models, particularly generative models, often alter the original text during prompt engineering, automatically adding, removing, or modifying Chinese characters. This leads to inconsistencies between the generated output and the original text, posing challenges for maintaining textual fidelity.

In EvaHan2024, numerous instances of such discrepancies were observed, where the model, while performing punctuation restoration, simultaneously modified the original sentence, resulting in unintended textual differences (Jin and Chen, 2008). This issue has persisted in the current evaluation, indicating that further attention is required in analyzing model outputs. To ensure that the generated results remain faithful to the original text, post-processing mechanisms should be incorporated into the workflow. Such mechanisms would help correct unintended modifications and restore textual accuracy, ensuring greater consistency between the model's output and the original input.

In this evaluation, most teams encountered issues with character omission and redundancy. The majority of differences of Chinese characters between the submitted results and the test set are around 1% to 2%, with the largest deviation reaching 8%. Although algorithms were employed in this evaluation to rectify the problems of character omission and redundancy in the submissions, teams still struggled to achieve high scores. Hence, to solve the issues of character omission and addition over-generated by large language models, post-processing is needed for the text consis-

tancy. Another way is to constrain the generated characters during model output generation to maintain consistency with the original text.

6.4 Teams' Approaches

In this evaluation, several teams adopted unique approaches to address the challenges of ancient Chinese NER, achieving notable improvements. Among them, the RUC team, which achieved the highest performance in this assessment, employed a combination of the GujiRoBERTa pre-trained model and the W2NER word-pair relation prediction framework. By leveraging BiLSTM and convolutional layers for feature extraction, along with five-fold cross-validation and ensemble learning, they significantly enhanced the effectiveness of ancient Chinese NER. Their method demonstrated outstanding results on the EvaHan2025 dataset, as shown in Table 9.

Looking at the overall evaluation results, most teams outperformed the baseline model. A comparative analysis reveals that, in addition to the adoption of innovative algorithms by some teams, the primary factor contributing to the improvement is the superior performance of GujiRoBERTa over SikuRoBERTa, which was used in this evaluation.

Moreover, some teams used prompt engineering techniques of large language models. However, these methods yielded limited improvements in performance and resulted in greater modifications to the original text, making them less effective for this task.

7 Conclusions

EvaHan2025 focuses on Named Entity Recognition in Ancient Chinese texts, covering two distinct categories of documents and presenting a significant challenge. Despite the complexity, most participating teams successfully completed the task. In terms of performance across different text types, teams generally performed better on historical texts, while their results on medical texts were comparatively lower, though still surpassing the baseline model.

From a methodological perspective, the majority of teams trained three separate models for each test set, achieving commendable results. However, no team has yet proposed a comprehensive, unified model capable of handling all 12 categories of named entities effectively. Additionally, a comparison of different implementation strategies reveals that prompt engineering based on large language models has shown limited effectiveness, often leading to undesirable modifications to the original text.

In the future, we encourage teams to explore deeper and more innovative approaches. Whether through small, domain-adaptive models or comprehensive frameworks leveraging large language models, we hope to see more efficient and accurate NER solutions for ancient Chinese, ultimately enabling high-performance, integrated recognition of diverse named entities across multiple categories. With the achievements of this shared task, we will move forward to the named entity relation recognition, named entity linking and related tasks in the coming years.

Acknowledgments

Thank the reviewers for their advices. Thank Yuyun Pan, Mengting Xu and Ye Gao for their data annotation and checking. This research was supported by General Program of the Ministry of Education of China Humanities and Social Sciences Fund (24A10319028), National Social Science Funds of China (21&ZD331) and Beijing Natural Science Foundation (7252253).

References

Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, 29:21–43.

Guangjin Jin and Xiao Chen. 2008. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Bin Li, Bolin Chang, Zhixing Xu, Minxuan Feng, Chao Xu, Weiguang Qu, Si Shen, and Dongbo Wang. 2024. Overview of EvaHan2024: The first international evaluation on Ancient Chinese sentence segmentation and punctuation. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 229–236, Torino, Italia. ELRA and ICCL.

Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang Qu, and Dongbo Wang. 2022. The first international Ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 135–140, Marseille, France. European Language Resources Association.

Li Dongmei, Luo Sisi, Zhang Xiaoping, and Xu Fu. 2022. Review on named entity recognition. *Journal of Frontiers of Computer Science & Technology*, 16(9):1954–1968.

Liu Liu and Wang Dongbo. 2018. A review on named entity recognition. *Journal of the China Society for Scientific and Technical Information*, 37(3):329–340.

L.F. Rau. 1991. Extracting company names from text. In *Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application*, pages 29–32.

Dongbo Wang, Litao Lin, Zhixiao Zhao, Wenhao Ye, Kai Meng, Wenlong Sun, Lianzhen Zhao, Xue Zhao, Si Shen, Wei Zhang, and Bin Li. 2023. EvaHan2023: Overview of the first international Ancient Chinese translation bakeoff. In *Proceedings of ALT2023: Ancient Language Translation Workshop*, pages 1–14, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia. Association for Computational Linguistics.