Overview of EvaHan2024: The First International Evaluation on Ancient Chinese Sentence Segmentation and Punctuation

Bin Li[1, 2, 3] Bolin Chang[1, 2] Zhixing Xu[1, 2] Minxuan Feng[1, 2] Chao Xu[1, 2]
Weiguang Qu[4, 2] Si Shen[5] Dongbo Wang[6, 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. Faculty of Arts and Humanities, University of Macau, China
4. School of Computer and Electronic Information, Nanjing Normal University, China
5. School of Economics and Management, Nanjing University of Science and Technology, China
6. College of Information Management, Nanjing Agricultural University, China
E-mail: db.wang@njau.edu.cn

Abstract

Ancient Chinese texts have no sentence boundaries and punctuation. Adding modern Chinese punctuation to theses texts requires expertise, time and efforts. Automatic sentence segmentation and punctuation is considered as a basic task for Ancient Chinese processing, but there is no shared task to evaluate the performances of different systems. This paper presents the results of the first ancient Chinese sentence segmentation and punctuation bakeoff, which is held at the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) 2024. The contest uses metrics for detailed evaluations of 4 genres of unpublished texts with 11 punctuation types. Six teams submitted 32 running results. In the closed modality, the participants are only allowed to use the training data, the highest obtained F1 scores are respectively 88.47% and 75.29% in sentence segmentation and sentence punctuation. The perfermances on the unseen data is 10 percent lower than the published common data, which means there is still space for further improvement. The large language models outperform the traditional models, but LLM changes the original characters around 1-2%, due to over-generation. Thus, post-processing is needed to keep the text consistancy.

Keywords: Ancient Chinese, Sentence Segmentation, Sentence Punctuation, Evaluation

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 understanding of these fundamental tasks (Li et al., 2022). The following year, the series moved to Macau, 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). In 2024, EvaHan is set to pioneer a new frontier with its first campaign specifically devoted to the evaluation of Ancient Chinese Sentence Segmentation and Punctuation, aiming to address a critical and yet under-explored area in the processing of classical texts.

In the natural language processing (NLP) tasks like speech to text recognition and chat text punctuation, texts often lack correct or appropriate sentence boundaries and punctuation (Nagy et al., 2021), a situation that increases the complexity of processing and reduces efficiency (Jones et al., 2003; Tündik et al., 2018). To enhance subsequent task processing, it is essential to add correct sentence boundaries and punctuation to these texts (Peitz et al., 2011). Addressing this, recent research has explored using large language models for automatically punctuating text in tasks such as text analysis and speech processing (Kolár and Lamel, 2012; González-Docasal et al., 2021; Bakare et al., 2023). Given the critical role of punctuation in text interpretation, comprehensive evaluations have been conducted to assess the effectiveness of automatic punctuation in NLP tasks (Meister et al., 2023). These evaluations have developed scientific indicators for texts in English and other languages, forming a complete and robust evaluation system.

Ancient Chinese also has no sentence boundaries and punctuation, making it quite hard to read (Lyu et al., 1983). Nowadays, in most republished ancient Chinese books punctuation is added manually by language experts. Here is an example of ancient Chinese.

(1)	亟	請	於	武公	公	弗	許
	repeatedly	request	to	Wugong	Wugong	not	accept

(Wu Jiang) repeatedly requested Wugong, but he refused.

Table 1 shows the sentence boundaries and punctuation added to Exp 1.

Raw Text	亟請於武公公弗許
+Sentence Segmentation	亟請於武公 公弗許
+Sentence Punctuation	亟請於武公，公弗許。

Table 1: Example of adding sentence segmentation and punctuation.

LT4HALA 2024@LREC-COLING-2024, pages 229–236
25 May, 2024. © 2024 ELRA Language Resources Association: CC BY-NC 4.0

With the establishment of the modern Chinese punctuation system, important texts of ancient books republished nowadays all include punctuation, which are much easier to read. But this work requires experts with great language knowledge of ancient Chinese. For example, a scholar usually needs several months to finish one book with around 200,000 characters. The great costs of time, funds and efforts place constraints on republication of these texts. And there is still a huge number of ancient books need to be processed. But most ancient books do not have that great value to be republished in paper books. The electronic texts could be automatically processed for many NLP tasks and applications, such as knowledge mining, Q&A, and machine translation (Sommerschield et al., 2023).

Therefore, automatic sentence segmentation and punctuation in ancient Chinese are fundamental tasks for compiling and publishing ancient books as well as ancient Chinese information processing, laying the foundation for subsequent tasks (Su et al., 2021). In recent years, research on sentence segmentation and punctuation in ancient Chinese have achieved good results (Chen et al., 2007; Huang et al., 2008; Shi et al., 2019; Yu et al., 2019; Cheng et al., 2020; Hong et al., 2021; Hu et al., 2021; Yuan et al., 2022), yet encountering some challenges.

Firstly, the number of types of punctuation used in existed automatic annotation systems vary from the basic 4 punctuation to 8. As a result, it is not easy to judge or compare the performances of the systems. Secondly, sentence segmentation and punctuation are usually conducted in a pipeline. Sentence segmentation errors will easily spread to the punctuation process. Thirdly, the evaluation paradigm for sentence segmentation and punctuation were not fully set up. The data set used for sentence segmentation and punctuation was disorganized, potentially due to the integration of test sets with training sets in the pre-training of large language models. And in calculating model scores, most studies rely on character-based assessments rather than punctuation-based assessments. Sentence segmentation and punctuation in ancient Chinese necessitate an evaluation task to address these challenges, to standardize irregular processes, and to provide a benchmark.

EvaHan2024 aims to give a good evaluation metric for this joint task and to answer three main questions:

- How can modern punctuation be integrated into ancient texts that lack sentence boundary and punctuation?
- Could the methodology of large language models facilitate processing ancient Chinese information?
- To ensure the integrity of the evaluation process, particularly given that large language models are trained on extensive collections of ancient Chinese texts, what strategies can be employed to prevent the overlap of the test corpus with the training set?

EvaHan2024 is proposed as part of the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA), co-located with LREC 2024. Scorer and detailed guidelines are all available in our GitHub repository[1].

2. Task

EvaHan2024 consolidated the following two problems into a joint task:

- Sentence segmentation involves converting Chinese text into a sequence of sentences, with each sentence separated by a single space.
- Sentence punctuation, on the other hand, focuses on the accurate placement of appropriate punctuation marks at the end of each sentence to ensure clarity.

In this shared task, a sentence should be automatically parsed from raw text to punctuated text shown in Table 1. There are eleven types of punctuation involved in the evaluation, as shown in Table 2. The evaluation toolkit gives the scores on both sentence segmentation and punctuation. EvaHan2024 does not accept running results with sentence segmentation only.

Punctuation	Name
，	Comma
。	Period
、	Slight-pause
：	Colon
；	Semicolon
？	Question Mark
！	Exclamation Mark
“	Left Quote
”	Right Quote
《	Left Book Title Mark
》	Right Book Title Mark

Table 2: 11 Punctuation involved in the evaluation

3. Dataset

The training dataset of EvaHan2024 is extracted from the classic historical books *Siku Quanshu* (四库全书)[2], the test data is extracted from 4 unpublished books. The comparison dataset is the text from *Zuozhuan*[3]. All the data has been punctuated and proofread by experts of Ancient Chinese language.

3.1 Data Format

The dataset consists of two parts, a training dataset and two test datasets, as shown in Table 3. All the punctuation are annotated by following *General Rules for Punctuation* (2012) and *Academic Publishing Specification-Collation of Chinese Ancient Books* (2015). All texts are encoded in UTF-8 plain text files.

[1] https://circse.github.io/LT4HALA/2024/EvaHan

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

[3] https://catalog.ldc.upenn.edu/LDC2017T14

As there are no sentence boundaries in Chinese texts, the raw texts only contain Chinese characters. After manual annotation, sentence punctuation are added to the text. As shown in Table 1, each sentence is marked with punctuation.

Data Sets	Sources	# Char Tokens	# Punctuation Tokens
Train	*Siku Quanshu*	19,796,102	3,929,523
TestA	4 genres of texts	50,306	9,673
TestB	*Zuozhuan*	196,560	53,919

Table 3: Texts distributed as training/test data in EvaHan2024.

3.2 Training Data

The training data includes punctuated text sourced from *Siku Quanshu* (四库全书), the largest series of ancient Chinese books , assembled during the Qing Dynasty. *Siku Quanshu* comprises four volumes including Jing, Shi, Zi and Ji, approximately 997 million words in total.

3.3 Test Data

Test Data was supplied in its raw format, consisting of Chinese characters only. Gold data was released after the evaluation period.

There are two test datasets. Blind *TestA* is designed to see how a system performs on dissimilar data. *TestA* includes four genres, namely *Products in Local Chronicles* (方志物产), *County Annals* (县志), *Buddhist Sutra* (佛经) and *Academy Records* (书院志), as shown in Table 4. TestA was not publicaly released/published publicly before EvaHan. This is an important way to ensure that no test data has been used by training procedure, especially for the LLM pre-training.

Genres	# Char Tokens	# Punctuation Tokens
Products in Local Chronicles (方志物产)	6,578	1,982
County Annals (县志)	24,548	4,244
Buddhist Sutra (佛经)	9,854	1,957
Academy Records (书院志)	9,326	1,490

Table 4: Four genres of TestA

We also compiled up a comparison test set *TestB*, which is designed to see how a system performs on similar data from the training data. *TestB* is the text of *Zuozhuan (左传)*, an ancient Chinese work believed to date back to the Warring States Period (475-221 BC). Specifically, *Zuozhuan* is a commentary on the *Chunqui (春秋)*, a history of the Chinese Spring and Autumn period (770-476 BC). *TestB* is partially included in the training set, and it can be easily obtained from the web. But the teams are not allowed to use it as training data directly. There have been several papers reporting their performance on this data (Shi et al., 2010; Cheng 2020 et al., 2020). Its size is larger than testA, containing 196,560 characters and 53,919 punctuation.

As *Zuozhuan* is included in *Siku Quanshu*, utilized for pre-training large language models, *TestB* serves solely as a reference for comparison.

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 was 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 sentence segmentation and punctuation were evaluated by precision, recall and F1 score. In the scoring process, we assess the correctness of punctuation directly, rather than Chinese characters as done in previous researches. The final ranking was based on F1 score of auto punctuation.

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 *Train*, and *XunziALLM*[4], a large language model pretrained on a very large corpus of traditional Chinese collection, including *Siku Quanshu* (四库全书). 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 5. But each team has to state all the resources, data and models they use in each system in the final report.

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

Table 5: Limitations on the two modalities.

4.3 Procedure

Training data was released for download from January 20, 2024. Test data was released on March

[4] https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM

1, 2024, and results were due on 00:00 (UTC) March 8, 2024.

5. Participants and Results

5.1 Participants

A total of 17 teams registered for the task, with 6 of those teams ultimately submitting 32 entries. Table 6 presents the details of the participating teams. Notably, the majority of submissions were under the 'closed modality', with only one team opting for the 'open modality'. It is important to mention that 27 submissions were initially presented in incorrect formats, as indicated by the '+' symbol in Table 6. This issue, primarily attributed to the over-generation of language by large language models (LLM), was subsequently rectified by us to facilitate accurate evaluation.

ID	Name	Affiliation	TestA		TestB	
			C	O	C	O
1	BNU	Beijing Normal University	1^{+}	0	1^{+}	0
2	CT	China Telecom Corporation Ltd. AI Technology Company	1^{+}	1^{+}	1^{+}	1^{+}
3	MiDU	Beijing Midu Information Technology Co., Ltd.	7^{+}	0	7^{+}	0
4	NJU1	Nanjing University	1^{+}	0	3	0
5	NJU2	Nanjing University	1^{+}	0	1	0
6	SU	Soochow University	1^{+}	0	1	0

Table 6: Participating teams by Corpus and Modality (**C**losed and **O**pen). Files with "+" means that the LLM changes the original texts.

5.2 Results

Table 7-10 list the performance of the participating teams, arranged in descending order of the F1 scores for the sentence punctuation. The Precision, Recall and F1 score for Sentence Segmentation, are abbreviated as P_{seg}, R_{seg} and F_{seg}, respectively. Simliarly, for sentence punctuation, they are abbreviated as P_{punc}, R_{punc} and F_{punc}. We categorized the results submitted by the participants as *TestA* Closed, *TestA* Open, *TestB* Closed, and *TestB* Open. The results are ranked by the sentence punctuation scores. Most teams participated in the closed tests. It can be seen from the four tables that there is a high correlation between sentence segmentation and sentence punctuation.

For *TestA*, the highest F1 score of sentence punctuation is 75.29% in the closed modality. In the open modality, it is 72.12%.

The scores of sentence segmentation are much higher. CT scores 88.86% and 87.93% in the closed and open modality. It is remarkble that MiDU scores 88.47% in the closed modality, with a slightly higher score 75.29% for sentence punctuation.

For TestB, which is designed to see how the systems perform on similar data as the training set, the scores have all increased by approximately 5 to 10 points. NJU2 scores 82.43% in TestB, ranking the first place in the close modality. But they submit no result in the open modality, and this score is even higher than their performance on TestA.

Team	P_{seg}	R_{seg}	F_{seg}	P_{punc}	R_{punc}	F_{punc}
MiDU	91.05	86.04	88.47	78.81	72.07	**75.29**
SU	89.84	84.70	87.19	75.88	69.71	72.67
CT	91.11	86.72	**88.86**	74.34	68.49	71.30
NJU2	90.80	76.34	82.94	77.75	63.85	70.12
NJU1	90.93	75.57	82.54	74.15	60.14	66.41
BNU	90.93	71.61	80.12	73.83	56.92	64.28

Table 7: TestA closed modality (%)

Team	P_{seg}	R_{seg}	F_{seg}	P_{punc}	R_{punc}	F_{punc}
CT	90.78	85.24	87.93	75.64	68.92	72.12

Table 8: TestA open modality (%)

Team	P_{seg}	R_{seg}	F_{seg}	P_{punc}	R_{punc}	F_{punc}
NJU2	95.98	90.54	93.18	85.08	79.93	**82.43**
CT	96.32	91.46	**93.83**	85.99	79.10	82.40
SU	94.64	91.93	93.27	82.93	78.96	80.89
MiDU	95.05	90.05	92.48	82.92	77.30	80.01
NJU1	95.38	89.68	92.44	80.44	75.67	77.98
BNU	95.25	88.15	91.57	79.06	73.66	76.26

Table 9: TestB (for comparison only) in closed modality (%)

Team	P_{seg}	R_{seg}	F_{seg}	P_{punc}	R_{punc}	F_{punc}
CT	94.73	89.21	91.89	82.91	74.94	78.73

Table 10: TestB (for comparison only) in open modality (%)

5.3 Baselines

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

5.3.1 Sentence Segmentation

The baseline for ancient Chinese sentence segmentation was constructed by *XunziALLM* (Xunzi-Qianwen-7B-CHAT) model, as shown in Table 11.

Testing Set	P_{seg}	R_{seg}	F_{seg}
TestA	90.53	66.12	76.42
TestB	95.28	87.17	91.04

Table 11: Sentence segmentation baselines (%)

The sentence segmentation scores of most teams exceed the baselines in TestA and TestB. The best scores outperform the baselines by around 10 points as shown in Table 12.

Testing Set	P_{seg}	R_{seg}	F_{seg}
TestA	91.11(+0.58)	86.72(+20.6)	88.86(+12.44)
TestB	96.32(+1.04)	91.46(+4.76)	93.83(+2.79)

Table 12: The promotion to the baselines of sentence segmentation (%)

5.3.2 Sentence Punctuation

The baseline for ancient Chinese sentence segmentation was constructed by *XunziALLM* model, as shown in Table 13.

Testing Set	P_{punc}	R_{punc}	F_{punc}
TestA	73.52	52.22	61.06
TestB	79.25	72.09	75.50

Table 13: Sentence punctuation baselines (%)

The sentence punctuation scores of most teams exceed the baselines in TestA and TestB. The best scores outperform the baselines by around 10 points as shown in Table 14.

Testing Set	P_{punc}	R_{punc}	F_{punc}
TestA	78.81(+5.29)	72.07(+19.85)	75.29(+14.23)
TestB	85.08(+6.74)	79.93(+7.84)	82.43(+6.93)

Table 14: The promotion to the baselines of sentence punctuation

5.4 Error Analysis

By analyzing the errors made by each team's system, we are able to observe different performances across different genres of texts and different punctuation.

5.4.1 Genres

Table 15 lists the F1 scores of teams in sentence segmentation and punctuation of texts in four genres. It becomes evident that most teams excelled in sentence segmentation and punctuation accuracy with *Products*, followed by county annals, and then academy records, while performance was notably lower with Buddhist sutra. The divergent performance across these four genres are examined as follows.

Firstly, the training set predominantly comprises data from genres such as county annals and academy records, with minimal representation from *Buddhist sutra*. Consequently, teams achieved markedly higher scores in county annals and academy records compared to *Buddhist sutra*, owing to the disparity in data within the training set.

Secondly, most teams gain the highest scores on *Products* data, despite its limited occurence in the training set. This is caused by the prevalence of slight-pauses and commas in *Products* data, typically occurring within lists of words devoid of complex vocabulary or syntactic structures. Example 2 is an example of *Products* data with many slight-pauses. Consequently, the model could achieve superior results on *Products* data through straightforward judgments.

(2) 打鐵鳥、黎母雀、紅頭鶯、鵓鴿、喜鵲、麻雀、山呼、鸚鵡、鴝鵒、秦吉了、五色雀、雉雞、烏、黃鶯、剪刀雀、鷓鴣、鳩、百舌、鵪鶉、杜鵑、畫眉、啄木、火雞、山雞、鴟鴞、蓑衣鶴、水鴨、白臉雞、鷺鷥、青莊、鶺鴒、翡翠、鵜鶘、鸂鶒、鴛鴦、割雀、鷗、海鵝、水鷹、海鳥、鶴、火烏、烏須、天鵝、知風、水晶、飛魚鳥、檳榔燕、華雞。

Team	Products		County		Buddhist		Academy	
	F_{seg}	F_{punc}	F_{seg}	F_{punc}	F_{seg}	F_{punc}	F_{seg}	F_{punc}
BNU	80.36	64.66	85.47	67.78	61.42	47.34	84.47	71.91
CT	**93.58**	**82.20**	**89.46**	73.91	83.44	50.47	87.96	75.6
MiDU	91.66	81.63	88.28	73.21	**85.43**	**71.80**	88.87	**77.43**
NJU1	88.23	73.04	83.23	64.95	74.04	58.95	83.67	70.97
NJU2	75.96	63.73	87.17	**74.10**	76.78	62.53	86.07	75.25
SU	91.38	81.85	88.13	72.13	79.01	61.91	**89.09**	75.46

Table 15: F1 scores for sentence segmentation and punctuation of texts in four genres (%)

5.4.2 Punctuation of Different Types

Table 16 lists the quantity of annotations and corresponding scores for different punctuation marks in the highest-scoring *TestA* submissions by MiDU. In Table 16, *TestA* (gold) means the number of gold punctuation in *TestA*. Machine (Total) means the total number of punctuation tagged by the MIDU's system running on *TestA*. Machine (Correct) means the number of correct punctuation tagged by MIDU's system. It is evident that comma exhibits the highest performance, while double quotation marks and book title punctuation perform less satisfactorily. There are three main issues with the system's performance in punctuation.

Puncs	P (%)	R (%)	F (%)	TestA (gold)	Machine (Correct)	Machine (Total)
、	92.34	71.24	80.43	1,269	904	979
，	77.34	79.23	78.27	4,949	3,921	5,070
。	76.38	76.67	76.52	2,332	1,788	2,341
？	77.5	70.45	73.81	88	62	80
！	93.33	48.28	63.64	29	14	15
；	76.92	45.98	57.55	87	40	52
：	77.12	44.36	56.32	266	118	153
《	87.72	27.78	42.19	180	50	57
》	82.46	26.11	39.66	180	47	57
“	66.67	10.07	17.5	139	14	21
”	63.16	8.82	15.48	136	12	19

Table 16: Punctuation scores by MIDU

First, the number of samples in the training set affects the effectiveness of punctuation annotation. Table 17 shows the distribution of punctuation marks in the training set. In conjunction with Table 16, it can be observed that punctuation marks with better annotation performance, such as commas, are more numerous in the training set, whereas punctuation marks with poorer performance ,such as book title marks, are less frequent. Therefore, to further improve the model’s performance, it would be advisable to select different corpora when creating the training set, to adjust the distribution consistency of punctuation marks within the training set.

Puncs	Count
，	1,879,220
。	954,948
：	163,968
、	126,394
“	120,769
”	119,407
？	73,067
《	60,302
》	60,256
；	55,256
！	45,623

Table 17: The distribution of punctuation marks in the training set

Second, the genres also affects the effectiveness of punctuation annotation. Despite the relatively sparse presence of commas in the training set, their strong performance can be attributed to the abundance of commas and periods in the text of *Products* (物产), which makes the annotation poccess easier and more accurate.

Thirdly, the issue of pairing exists in the use of paired punctuation marks. Among the eleven types of punctuation marks, double quotes and book title marks are different from others in that they appear in pairs. These paired punctuation marks have some specific requirements in annotation : after a left quote, there must be a right quote, and not another left quote, and the number of left and right quotes must be the same. However, according to the data in Table 16, the number of left quotes annotated by machine does not equal the number of right quotes. Although 21 left quotes were annotated, only 19 right quotes appeared. This indicates that post-processing can be used to further improve the performance of the systems.

6. Discussion

6.1 Consistency in Paired Punctuation Marks

In the evaluation of various punctuation types within the submissions, a notable inconsistency was observed in the usage of inherently paired punctuation marks, such as double quote marks and book title marks. This inconsistency was particularly evident in one team's submission, where a significant imbalance was recorded: the frequency of left quote marks was nearly fivefold that of right quote marks. Although numerous teams have employed specialized post-processing techniques to address character omission and addition issues common in large language models, these efforts appear to have insufficiently accounted for the nuances of Chinese punctuation. Moreover, a critical oversight in these submissions is the lack of consistency checks for paired punctuation marks. Such checks are essential for ensuring punctuation accuracy, especially in the context of complex language structures like those found in Chinese.

6.2 Implementation Strategy for Book Title Marks

The low performance in handling book title marks, as observed in this evaluation, stems from two main issues: inconsistent handling across different cases, and the approach adopted for processing quote marks. Book title marks, which are used to denote book titles, chapter names, and similar entities, warrant a specific treatment due to their distinct significance. In fact, the annotation of these marks could be effectively treated as a task of named entity recognition, primarily focusing on book titles. Previous studies have approached book title marks as individual named entities, yielding some successful outcomes. However, during this evaluation, it became evident that participating teams did not develop specialized solutions for book title marks. Instead, they handled them as generic punctuation marks and failed to observe their specific function and importance.

6.3 Character Discrepancies Due to Large Language Models

Large language models, particularly generative ones, often alter the original text during prompt engineering, automatically adding or removing Chinese characters, leading to discrepancies between the output and the original text. 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 consistancy. Another way is to constrain the generated characters during model output generation to maintain consistency with the original text.

7. Conclusions

EvaHan2024 marks a pioneering endeavor in the field of ancient Chinese sentence segmentation and punctuation. The best system of this bakeoff, developed by MiDU, notably outperformed the majority of its counterparts. The deployment of large language models has indeed elevated performance metrics in processing ancient Chinese texts. The test sets have a wide coverage and one was implemented as a blind test, therefore, the effectiveness of sentence segmentation and punctuation is more challenging than expected, leaving ample room for improvement. It is noteworthy that even advanced language models are not immune to issues such as character omission and excessive generation. Therefore, it is imperative for participating teams to actively engage with and address these complexities. In the future, the next iteration of EvaHan should broaden its scope to encompass a wider array of genres and cross-temporal corpora. This expansion is anticipated to foster improvements in handling a more diverse set of data.

8. Acknowledgements

Thank the reviewers for their advices. Thank Yongji Wang, Jingxuan Xi, Pengxiu Lu, Ruijia Yang, and Han Xiao for their data annotation and checking. Thank GULIAN (Beijing) Media TECH CO.,LTD for their data support. This research was supported by National Social Science Funds of China (21&ZD331), and National Language Commission Project of China (YB145-41).

9. References

Bakare, A. M., Anbananthen, K. S. M., Muthaiyah, S., Krishnan, J., and Kannan, S. (2023). Punctuation Restoration with Transformer Model on Social Media Data. *Applied Sciences*, 13(3), 1685.

Chen, T., Chen, R., Pan, L., Li, H., and Yu, Z. (2007). Archaic Chinese Punctuating Sentences Based on Context N-gram Model. *Computer Engineering* (03), 192-193+196.

Cheng, N., Li, B., Ge, S., Hao, X., and Feng, M. (2020). A Joint Model of Automatic Sentence Segmentation and Lexical Analysis for Ancient ChineseBased on BiLSTM-CRF Model. *Journal of Chinese Information Processing*, 34(04), 1-9.

General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China. (2012). *General Rules for Punctuation* (GB/T 15834-2011). Beijing: Standards Press of China.

Hong, T., Cheng, R., Liu, S., and Fang, K. (2021). An Automatic Punctuation Method Based On the Transformer Model. *Digital Humanities* (02), 111-122.

Hou, H. and Huang, J. (2008). On Sentence Segmentation and Punctuation Model for Ancient Books on Agriculture. *Journal of Chinese Information Processing* (04), 31-38.

Hu, R., Li, S., and Zhu, Y. (2021). Knowledge Representation and Sentence Segmentation of Ancient Chinese Based on DeepLanguage Models. *Journal of Chinese Information Processing*, 35(04), 8-15.

Jin, G. and Chen., X. (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*.

Jones, D. A., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D. A., and Zissman, M. (2003). Measuring the Readability of Automatic Speech-To-Text Transcripts. *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 1585–1588.

Kolár, J., and Lamel, L. (2012). Development and evaluation of automatic punctuation for french and english speech-to-text. *Interspeech*, 1376-1379.

Li, B., Yuan, Y., Lu, J., Feng, M., Xu, C., Qu, W., and Wang, D. (2022). The First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff: Overview of the Evahan 2022 Evaluation Campaign. *Proceedings of the second workshop on language technologies for historical and ancient languages*, 135-140.

Lyu S. (1983). The First Step in Organizing Ancient Texts. *China Publishing Journal*, 71(4), 44-50.

Meister, A., Novikov, M., Karpov, N., Bakhturina, E., Lavrukhin, V., and Ginsburg, B. (2023). *LibriSpeech-PC: Benchmark for Evaluation of Punctuation and Capitalization Capabilities of end-to-end ASR Models. Proceddings of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*: 1-7.

Nagy, A., Bial, B., and Ács, J. (2021). Automatic punctuation restoration with BERT models. arXiv:2101.07343v1.

Peitz, S., Freitag, M., Mauser, A., and Ney, H. (2011). Modeling Punctuation Prediction as Machine Translation. *Proceddings of the International Workshop on Spoken Language Translation*, 238–245.

Shi, X., Shi, X., Shi, X., Shi, X., and Song, Y. (2019). A Method and Implementation of Automatic Punctuation. *Journal of Digital Archives and Digital Humanities* (3), 1-19.

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., and Freitas, N. D. (2023). Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, 1-45.

Su, Q., Hu, R., Zhu, Y., Yan, C., and Wang, J. (2021). Key Technologies for Digitization of Ancient Chinese Books. *Digital Humanities Research* (03), 83-88.

The State Administration of Press, Publication, Radio, Film and Television of the People's Republic of China. (2015). *Academic Publishing Specification-Collation of Chinese Ancient Books* (CY/T 124-2015). Beijing: Standards Press of China.

Tündik, M. Á., Szaszák, G., Gosztolya, G., and Beke, A. (2018). User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning. *Interspeech* 2018, 2628–2632.

Wang, D., Lin, L., Zhao, Z., Ye, W., Meng, K., Sun, W., Zhao, L., Zhao, X., Shen, S., Zhang, W., and Li, B. (2023). EvaHan2023: Overview of the First International Ancient Chinese Translation Bakeoff. *Proceedings of ALT2023: Ancient Language Translation Workshop*, 1-14.

Yu, J., Wei, Y., and Zhang, Y. (2019). Automatic Ancient Chinese Texts Segmentation Based on BERT. *Journal of Chinese Information Processing* (11), 57-63.

Yuan, Y., Li, B., Feng, M., He, S., and Wang, D. (2022). A Joint Model of Automatic Sentence Segmentation and Punctuation for Ancient Classical TextsBased on Deep Learning. *Library and Information Service*, 66(22), 134-141.