Evaluating Evaluation Metrics for Ancient Chinese to English Machine Translation

Eric Bennett[1,2], HyoJung Han[1], Xinchen Yang[1], Andrew Schonebaum[2], Marine Carpuat[1]
[1]Dept. of Computer Science, University of Maryland, College Park
[2]Dept. of East Asian Languages and Cultures, University of Maryland, College Park
ebenne92@umd.edu, hjhan@cs.umd.edu, xcyang@cs.umd.edu,
schone@umd.edu, marine@cs.umd.edu

Abstract

Evaluation metrics are an important driver of progress in Machine Translation (MT), but they have been primarily validated on high-resource modern languages. In this paper, we conduct an empirical evaluation of metrics commonly used to evaluate MT from Ancient Chinese into English. Using LLMs, we construct a contrastive test set, pairing high-quality MT and purposefully flawed MT of the same Pre-Qin texts. We then evaluate the ability of each metric to discriminate between accurate and flawed translations.

1 Introduction

Large Language Models (LLM) make it possible to translate between languages in a zero-shot fashion. This makes it possible for English readers to access previously untranslated texts in ancient languages such as Ancient Chinese (Jin et al., 2023) or Latin (Volk et al., 2024). However, how can we determine how good these translations are? For our language of interest, Ancient Chinese, machine translation (MT) research has relied on standard reference-based metrics to assess translation quality, but these metrics have not been validated specifically for this language.

Ancient Chinese [1] presents a unique challenge in translation to English due to the language's laconic and epigrammatic nature, as well as the relatively limited resources available compared to other languages. There are numerous English translations of the most famous Ancient Chinese texts, including *Tao Te Ching* (Campbell, 2022), *Analects* (Jin et al., 2023), and *Dream of the Red Chamber* (Kong, 2022), but a large majority of texts remain inaccessible to English readers (Fordham, 2021). When translating Ancient Chinese into English, many Chinese characters have multiple meanings depending on their usage in a sentence, requiring disambiguation in the translation process (Zou, 2016). The large amount of idioms and symbolic language also makes translation difficult, along with a lack of sentence boundaries or punctuation, explicit plurals, or conjunctions, making it a uniquely difficult translation problem. (Li et al., 2024). While the advent of LLMs has led to improvements in MT quality for Ancient Chinese to English translation, current models still lag behind human translators. (Jin et al., 2023).

The complexity of translating from Ancient Chinese to English is reflected in the complexity of evaluation. Translations may capture the meaning of a sentence very well, while having very different wording from another valid English translation. This might be problematic when evaluating with metrics such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), which measure the word or character n-gram overlap between the MT output and a human-written reference translation. Neural metrics based on fine-tuning LLMs (Guerreiro et al., 2024; Rei et al., 2020; Juraska et al., 2023) have been found to correlate better with human ratings of translation quality for modern language pairs evaluated at the Conference on Machine Translation, including English-German and Japanese-Chinese (Freitag et al., 2024), but they have not been evaluated on translation from Ancient Chinese to English.

In this paper, we ask how well existing MT metrics are able to discriminate between 'good' and 'bad' English translations of Ancient Chinese texts. Building on meta-evaluation methods used for modern languages (Karpinska et al., 2022; Edunov et al., 2020), we address this question using a contrastive test set created by prompting an LLM for 'good' and 'bad' translations of the same Chinese inputs. After validating that the 'bad' translations are rated as worse than the 'good' translations by human

[1]The term Ancient Chinese encapsulates thousands of years of linguistic development (Chang et al., 2021). Our experiments use a Pre-Qin dataset from before the establishment of the Qin Dynasty in 221 BCE.

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 71–76
May 4, 2025.

judges, we use this set to evaluate the ability of standard MT evaluation metrics to discriminate between 'good' and 'bad' translations.

2 Test Set Construction

2.1 Data Collection

The dataset used for this experiment is a collection of texts from the Pre-Qin period (prior to the establishment of the Qin Dynasty in 221 BCE) acquired from Dongbo Wang's team at Nanjing Agricultural University (Li et al., 2024). The format of the Pre-Qin dataset is a collection of Ancient Chinese source texts paired with single human English reference translations.

We cleaned the data for this experiment by removing pairs with the following properties:

1. The source text contains English.
2. The source or target length is greater than one standard deviation from the mean (>61 characters), to simplify human validation.
3. Being a duplicated source text.
4. The text contains portions of the Tao Te Ching, as the high interpretability of the document could interfere with this evaluation.[2]

In total, from the original dataset of 23,686 source-reference pairs 6,794 were deleted in the data cleaning process, resulting in a set of 16,892 source-reference pairs for analysis. The results show insights from both the entire cleaned Pre-Qin dataset, and a 500 entry human validated sample drawn randomly from the Pre-Qin dataset (Table 1).

2.2 Synthetic Translations Generation

We used OpenAI's gpt-4o model (Hurst et al., 2024) to generate a 'good' and a 'bad' translation for each of the source texts. We used the following prompts for the 'good' and 'bad' outputs, respectively:

- "Translate the Ancient Chinese text into English. Respond with the translation only."
- "Translate the Ancient Chinese text into English incorrectly, deliberately introducing disambiguation errors, accuracy errors, and tense errors in the text. Respond with the translation only."

The error types listed in the 'bad' translation prompt were chosen based on common errors identified in Chinese to English translations (Freitag et al., 2021), and tense error was drawn from the lack of tense in Ancient Chinese.

Here is a randomly selected example from the evaluation dataset resulting from this process:

Source:
鮮卑寇酒泉；種衆日多，緣邊莫不被毒。

Reference translation:
The Xianbi raided Jiuquan. The numbers of their people increased day by day, and there was no region of the border country which did not suffer from them.

'Good' translation:
The Xianbei raided Jiuquan; their numbers grew daily, and the border regions suffered widespread harm.

'Bad' translation:
The Xianbei invaded Qiuquan; the people often multiply their seeds, along the edges they refuse to receive poison.

2.3 Human Validation

We asked human judges to validate the LLM-generated translations. A sample of 500 entries was randomly selected from the cleaned dataset, and given to two human evaluators, one being an expert with extensive experience in Classical Chinese to English translation, and one being a native Chinese speaker with an intermediate level of experience with Classical Chinese. 100 entries were randomly selected from the sample as a cross-validation set to ensure coherence between the validators, and each validator was given 200 unique entries to complete the 500 entry sample. The composition of the sample is shown in Table 1.

	# Entries	# Src Char	# Ref Char	# Ref Words
Sample	500	9,641	70,902	12,916
Pre-Qin	16,892	332,355	2,463,235	448,386

Table 1: Dataset summary

Each validator was given access to the source and reference for an entry, and asked to compare the quality of two unlabelled machine translations A and B by selecting one of 3 options: "A is better than B", "B is better than A", or "too hard to tell". The order in which the 'good' and 'bad' translations were provided was randomly assigned. Annotators did not receive explicit guidelines defining what makes a translation better, and were simply asked to rate based on their own best judgment (Vilar et al., 2007).

[2]Tao Te Ching is one of the most translated texts in the world, with over 2,052 recognized translations in 92 languages. (Tadd, 2022)

The Cohen's Kappa score was 0.78 on the doubly annotated subset, indicating a high strength of agreement. The two validators both chose the 'good' translation as higher quality in 88/100 entries. In entries where both validators decided on one of the translations (neither validator chose the "too hard to tell" option), there was an 88/90 (97.78%) accuracy, and there were no cases where both validators agreed that the 'bad' translation was better. For the compiled validation dataset of 500 entries, when differences between the two evaluators were present, the more expert evaluator response was chosen. Overall, the human validators selected the 'good' translation as higher quality in 471/500 entries (94.2%).

3 Metric Selection

When deciding which metrics to test, the first consideration was the metrics used in past papers regarding Ancient Chinese MT. The results of an analysis of 5 recent papers related to Ancient Chinese machine translation is located in Table 2. The "Other" metrics include Ancient Chinese LLM evaluation metrics not related to machine translation in Zhang and Li (2023) as well as LMS (Levenshtein-distance-based Morphological Similarity) and ESS (Embedding Semantic Similarity) for evaluation as proposed in Wang et al. (2023). With this in mind, SacreBLEU (Post, 2018) and ChrF++ were selected for testing.

Previous Works	BLEU	ChrF++	Neural	Other
Jin et al. (2023)	Multi ref	×	×	✓
Wang et al. (2023)	Single ref	✓	×	×
Nehrdich et al. (2023)	Single ref	✓	×	×
Chang et al. (2021)	Single ref	×	×	×
Zhang and Li (2023)	×	×	×	✓

Table 2: Evaluation metrics for Ancient Chinese MT in previous literature.

Furthermore, we decided to test the current state-of-the-art neural metrics for MT evaluation (Freitag et al., 2024) as well, despite them not being trained specifically on Ancient Chinese. From Google, metricx-24-hybrid-xl-v2p6 (Juraska et al., 2024) and metricx-23-xl-v2p0 (Juraska et al., 2023) were chosen. Both metrics are based on the mT5 encoder-decoder language model (Xue et al., 2021). MetricX-23 is finetuned using two stages of training, on direct assessment (DA) followed by MQM training data, as well as synthetic training data. MetricX-24 significantly expands the usage of synthetic data, and mixes DA and MQM data in the second training stage. MetricX-24 Hybrid allows for reference-based or reference-free evaluation in a unified model (in this experiment a reference is given) and had the highest correlation with human evaluation in WMT-24 with the exception of Meta-Metrics-MT (Anugraha et al., 2024).

Two COMET metrics were also chosen for analysis. XCOMET-XL (Guerreiro et al., 2024) is similar to MetricX-24 Hybrid in its ability to evaluate with or without a reference. It is based on the XLM-R XL encoder-decoder model (Conneau et al., 2020), and trained on DA data, followed by MQM data, and finally further high-quality MQM data. It also incorporates error-span detection in the training process, with the error-span detection function of the model sharing a common encoder with the sentence-level score function. COMET-WMT22 (Rei et al., 2022) is based on the XLM-R base model. It is trained primarily on DA data, followed by fine-tuning on z-normalized MQM scores.

For each of the selected metrics, we evaluated the two machine-translated hypotheses for each of the source entries. The provided single human reference translation was used as a single reference.

4 Results

To analyze the results of our evaluations using the chosen metrics, a difference score was calculated for each entry by subtracting the metric's score on the 'bad' translation from the score on the 'good' translation. A difference score of >0 represents a 'correct' prediction- that the generated 'good' translation was judged better than the 'bad' translation. A Wilcoxon signed-rank test was also performed for each metric to determine whether the ability of the metric to detect differences in scores is statistically significant. The performance of each metric, both on the entire 16,892 entry Pre-Qin dataset and the 500 entry human-validated sample, is described in Table 3, and Figure 1 compares distributions for each of the metrics in the human validated sample.

One notable performance from the evaluation is the following case, where all four neural metrics performed particularly poorly. The difference score for the evaluation fell within the bottom 10% for each metric, with the 'bad' translation being predicted as being higher quality than the 'good' translation by every metric except for MetricX-24 Hybrid despite the error of the direction 'left' being translated as 'right' in the 'bad' translation:

	Human Validated Sample						Pre-Qin Dataset					
metric	**% predicted correctly**	**mean**	**median**	**stdev**	**Wilcoxon Test Statistic**	**P-Value**	**% predicted correctly**	**mean**	**median**	**stdev**	**Wilcoxon Test Statistic**	**P-Value**
SacreBLEU	71.600	0.027	0.011	**0.064**	94892	9e-28	72.241	0.029	0.012	**0.066**	110817290	**0.0**
CHRF++	79.200	0.062	0.053	0.082	110333	1e-49	80.440	0.064	0.054	0.084	126219585	**0.0**
XCOMET-XL	88.000	0.175	0.170	0.150	119650	6e-70	88.048	0.168	0.156	0.149	136247392	**0.0**
COMET-WMT22	93.200	0.111	0.106	0.080	122582	4e-77	93.760	0.109	0.104	0.077	140339475	**0.0**
MetricX-24-XL	**95.800**	**0.230**	**0.223**	0.139	**124586**	**3e-82**	**95.803**	**0.226**	**0.223**	0.136	**141675253**	**0.0**
MetricX-23-XL	94.800	0.173	0.158	0.131	123238	9e-79	93.926	0.170	0.157	0.128	140164192	**0.0**

Table 3: Difference score metrics on validated sample and Pre-Qin dataset with Wilcoxon Test Statistic. For SacreBLEU and the two MetricX metrics scores were normalized between 0 and 1.

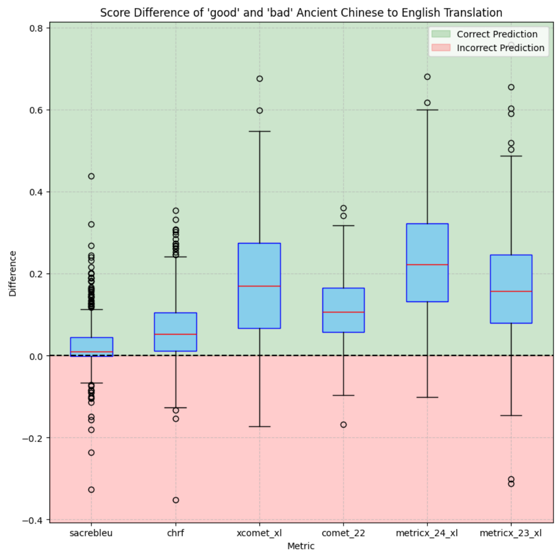

Figure 1: Box plot of difference scores. For SacreBLEU and the two MetricX metrics scores were normalized between 0 and 1.

Source:
有杕之杜: 有杕之杜、生于道左。
Reference translation:
You Di Zhi Du: There is a solitary russet pear tree,Growing on the left of the way.
'Good' translation:
A solitary tree in the woods: A solitary tree in the woods, growing by the roadside.
'Bad' translation:
There is a single pine tree: There is a single pine tree, growing on the right of the road.

Although all of the metrics were shown to have statistically significant success in the task of determining between the 'good' and 'bad' translations, some metrics performed with greater accuracy or more consistently. Commonly used metrics like BLEU and ChrF++ notably showed a lower standard deviation and therefore more consistency compared to newer metrics, with the exception of COMET-WMT22. While XCOMET-XL has a higher mean than COMET-WMT22, its higher variability results in a worse performance than the older model at predicting the 'good' translation. Furthermore, Figure 2 describes the correlation between metrics, showing that neural metrics tend to agree with each other more than with surface metrics, but still hold disagreements, particularly across families of models.

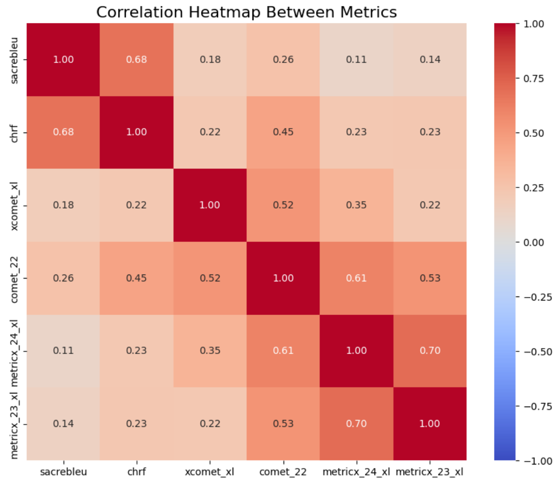

Figure 2: Correlation of difference scores between normalized metrics.

Overall, these results show that neural metrics are better able to discern 'good' and 'bad' translations than surface metrics, despite not being trained with translation quality ratings of MT from Ancient Chinese to English. Supervision from other MT tasks into English helps identify the problematic outputs in our test set. These results suggest future research on MT from Ancient Chinese would benefit from including neural metrics such as XCOMET-XL or MetricX-24 Hybrid to guide system development. At the same time, it would be useful to design metrics that target error categories known to be problematic for Ancient Chinese MT: the method we used here to generate contrastive synthetic translations could be extended to evaluate each metric's ability to detect specific error categories, and to provide training data for more targeted metrics.

Acknowledgments

We sincerely thank Professor Dongbo Wang and his team at Nanjing Agricultural University for providing access to the 中古代典籍跨言料 (Chinese Ancient Texts Cross-Language Corpus), which served as the source of Ancient Chinese to English translation data for this paper.

References

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. 2024. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, pages 459–469, Miami, Florida, USA. Association for Computational Linguistics.

Larry N. Campbell. 2022. *The Parallel Tao Te Ching: A Comparison of English Translations*, volume 1. Aftermath Enterprises LLC, Midland, TX.

Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, and Vera Demberg. 2021. Time-aware Ancient Chinese text translation and inference. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 1–6, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Carl Gene Fordham. 2021. English translations of chinese texts from the pre-qin through han period (2010–2020): Publishing trends and quality assurance.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

OpenAI: Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, and et al. 2024. Gpt-4o system card. *arXiv preprint*.

Kai Jin, Dan Zhao, and Wuying Liu. 2023. Morphological and semantic evaluation of Ancient Chinese machine translation. In *Proceedings of the Ancient Language Processing Workshop*, pages 96–102, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

L. Kong. 2022. Bonsall: The first full translation of the dream of the red chamber. *Advances in Literary Study*, 10(3):291–297.

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.

Sebastian Nehrdich, Marcus Bingenheimer, Justin Brody, and Kurt Keutzer. 2023. MITRA-zh: An efficient, open machine translation solution for buddhist Chinese. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 266–277, Tokyo, Japan. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Misha Tadd. 2022. *"Laozi" Yi Ben Zong Mu: Quan Qiu Lao Xue Yao Lan = the Complete Bibliography of Laozi Translations: A Global Laozegetics Reference*. Nan kai da xue chu ban she, Tianjin.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic. Association for Computational Linguistics.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 122–128, Torino, Italia. ELRA and ICCL.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yixuan Zhang and Haonan Li. 2023. Can large language model comprehend Ancient Chinese? a preliminary test on ACLUE. In *Proceedings of the Ancient Language Processing Workshop*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Siyu Zou. 2016. On ancient chinese translation. In *Proceedings of the 2016 6th International Conference on Management, Education, Information and Control (MEICI 2016)*, pages 675–678. Atlantis Press.