

PhiloGPT: A Philology-Oriented Large Language Model for Ancient Chinese Manuscripts with Dunhuang as Case Study

Yuqing Zhang^{1*}, Baoyi He^{1*}, Yihan Chen¹, Hangqi Li¹, Yue Han¹, Shengyu Zhang^{1,2†},
Huaiyong Dou¹, Junchi Yan^{3,4}, Zemin Liu¹, Yongquan Zhang¹, Fei Wu^{1,2,4†}

¹Zhejiang University, Hangzhou, China

²Shanghai Institute for Advanced Study of Zhejiang University, Shanghai, China

³Shanghai Jiao Tong University, Shanghai, China

⁴Shanghai AI Laboratory, Shanghai, China

{yuqingz7, 12321037, 3210100995, zjhanyue, 3210105819, sy_zhang, liu.zemin, wufei}@zju.edu.cn

douhyong@163.com, yanjunchi@sjtu.edu.cn, zdzyq@emb.zju.edu.cn

Abstract

Philology, the study of ancient manuscripts, demands years of professional training in extensive knowledge memorization and manual textual retrieval. Despite these requirements align closely with strengths of recent successful Large Language Models (LLMs), the scarcity of high-quality, specialized training data has hindered direct applications. To bridge this gap, we curated the *PhiloCorpus-ZH*, a rich collection of ancient Chinese texts spanning a millennium with 30 diverse topics, including firsthand folk copies. This corpus facilitated the development of *PhiloGPT*, the first LLM tailored for discovering ancient Chinese manuscripts. To effectively tackle complex philological tasks like restoration, attribution, and linguistic analysis, we introduced the *PhiloCoP* framework. Modeled on the analytical patterns of philologists, *PhiloCoP* enhances LLM's handling of historical linguistic peculiarities such as phonetic loans, polysemy, and syntactic inversions. We further integrated these tasks into the *PhiloBenchmark*, establishing a new standard for evaluating ancient Chinese LLMs addressing philology tasks. Deploying *PhiloGPT* in practical scenarios has enabled Dunhuang specialists to resolve philology tasks, such as identifying duplication of copied text and assisting archaeologists with text completion, demonstrating its potential in real-world applications.

1 Introduction

The dissemination of ancient human civilizations relied on various carriers, from inscriptions to manuscripts and the more recent printed texts (Assael et al., 2022). Manuscripts, particularly popular from the 4-th to the 14-th century due to advancements in papermaking and its affordability, encompassed a broad array of content including official documents and folk copies on economics, literature, science, and agriculture. These

manuscripts are invaluable for understanding ancient societies, offering insights into both the official narratives and the daily lives of the common people. However, interpreting these vast and topic-diverse ancient manuscripts which span over a millennium, presents a highly specialized challenge. It demands philologists to undergo years of professional training and engage in extensive literature reviews, manual transcription, and contextual analysis (Galambos, 2020).

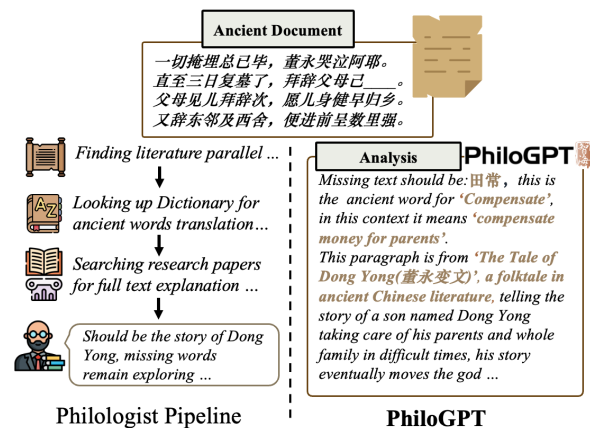


Figure 1: Illustration of PhiloGPT capabilities.

Aforementioned challenges render manuscript discovery a time-consuming and labor-intensive field of study. However, the requirements of philology align closely with the strengths of LLMs (Yin et al., 2023; Zhao et al., 2023). The recent success of LLMs in a variety of natural language processing (NLP) tasks across diverse fields has highlighted their effectiveness (Achiam et al., 2023; Xi et al., 2023; Anil et al., 2023). Nevertheless, the potential of LLMs to interpret ancient languages like Classical Chinese remains largely untapped (Sommerschild et al., 2023). This oversight is primarily due to three major constraints: (1) *Insufficient Training Corpus*. Despite the abundance of high-quality English, Chinese, and multilingual LLM training corpora (Minaee et al., 2024), there

*Equal Contribution. † Corresponding authors.

remains a significant gap in the availability of a large and diverse corpus for Classical Chinese. (2) *Linguistic Complexities of Classical Chinese*. The substantial differences in linguistic peculiarities between ancient and modern Chinese, such as words meaning shift and syntactic inversions, has complicated text interpretation. These phenomena demand a framework that integrates philological insights to fully exploit the capabilities of LLMs. (3) *Fragmentation in Ancient Language Research*. Earlier efforts often focused in isolated tasks (Yoo et al., 2022), leading to piecemeal approaches that lack an integration for comprehensive philological analysis. These fragmented efforts have restricted the ability to perform a holistic evaluation, limiting the depth of insights that can be gained.

To address these hurdles and fully leverage the capabilities of LLMs in Classical Chinese studies, we compiled *PhiloCorpus-ZH*, an extensive collection of ancient manuscripts. This pioneering archive systematically categorizes thousands of documents across 30 varied topics, including previously unorganized firsthand folk copies, enriching the understanding of ancient linguistic practices.

Recognizing the profound linguistic disparities between classical and modern Chinese, including grammatical evolution and syntactic variations, we developed the *PhiloCoP* framework. The essence of *PhiloCoP* lies in mimicking philologists internally interpreting ancient manuscripts, i.e., entity identification, context-implicit relation reasoning, and relation-aware transcription. Entity identification extracts, differentiates and categorizes entities based on their roles within the text. Context-implicit relation reasoning uncovers the underlying semantic connections and facilitates a form of cognitive mapping. Relation-aware transcription performs lexical standardization by adjusting for linguistic shifts and anomalies inherent in Classical Chinese. Eventually, *PhiloCoP* concludes with a synthesis prompt that requires the model to integrate its observations from the previous steps into a cohesive interpretation.

Furthermore, we propose *PhiloBenchmark* to evaluate the efficacy of LLMs from the perspective of philological research requirements. It comprises 9 targeted tasks closely related to ancient document analysis, including manuscript restoration, attribution, and philology-oriented QA, along with content reasoning. Building on all previous substantial resources, we developed *PhiloGPT*, the first LLM tailored to handle a variety of analyti-

cal tasks for Classical Chinese. Through extensive experiments, we demonstrate that *PhiloGPT*, in comparison to general LLMs, excels in the discovery and analysis of ancient documents, such as manuscript text pieces conjugation and historical reasoning. *PhiloGPT* is applying in Dunhuang manuscript research, demonstrating potential in tasks, such as distinguishing original and copied texts, and resolving debate among missing characters in manuscripts. The explored real-world applications of *PhiloGPT* showcase its versatility and value as an indispensable tool for scholars, accelerating discovery and fostering interdisciplinary collaboration in Classical Chinese studies.

The highlights of the paper are:

1. Diverse Original Ancient Chinese Corpus: We curated the *PhiloCorpus-ZH*, a collection of ancient Chinese manuscripts. These original documents, complemented by related modern research works, were systematically organized into four majors adhering to bibliographic principles.

2. Effective Reasoning Paradigm for Ancient Text: We introduce *PhiloCoP*, a framework that adheres to chain-of-thoughts for philologists, progressively navigating vocabulary and grammatical phenomena in classical Chinese. This approach demonstrates marked improvements across multiple downstream tasks.

3. Comprehensive Benchmark for Documentary Studies: We have compiled *PhiloBenchmark*, which includes 9 unique tasks to assess the performance of ancient Chinese LLMs specialized for manuscript discoveries. This benchmark addresses a wide range of research needs and skill requirements essential for philologists.

4. Real-world Applications with Dunhuang as Case Studies: *PhiloGPT* has been applied to specialized tasks tailored for Dunhuang scholars, including the restoration and distinction of manuscripts. This interdisciplinary case study shows our *PhiloGPT*'s further potential in assisting with philology discoveries.

2 Related Work

Domain-specific Large Language Models. Recent advancements in LLMs, particularly with the development of open-source frameworks, have catalyzed the development of many domain-specific LLM tailored for diverse downstream applications (Zhao et al., 2023). For instance, in the medical field, LLMs are now utilized for assistive di-

agnostics(Xiong et al., 2023), health-related question answering (QA)(Wang et al., 2023a), and psychological counseling(Chen et al., 2023). In the legal sector, specialized LLMs facilitate legal QA(Huang et al., 2023), generate legal documents(Cui et al., 2023a), and predict legal judgments(Yue et al., 2023). Further applications are seen in finance(Wu et al., 2023) and education(Xu et al., 2024), where LLMs process vast domain-specific datasets. Building these domain-specific models typically requires continuous pre-training on existing general LLMs, supplemented by supervised instruction tuning and task-specific alignment(Wei et al., 2021; Ouyang et al., 2022; Zhang et al., 2020). However, the scarcity of large-scale, original ancient text corpora and suitable task datasets has hindered the development of LLMs for ancient languages(Sommerschild et al., 2023). To address this gap, our work not only compiles a rich collection of ancient Chinese manuscripts, but also pioneers the training of the first LLM dedicated to ancient Chinese studies.

Language Models for Ancient Texts. Several noteworthy projects have successfully employed language models pre-trained on ancient language corpora for specialized tasks(Kang et al., 2021; Parker et al., 2019; Papavassileiou et al., 2023). For instance, (Assael et al., 2022) utilized ancient Greek inscriptions to support textual restoration, geographical and chronological attribution. Similarly, HUE’s work(Yoo et al., 2022) involved training BERT-like models on ancient Korean documents to address challenges such as chronological attribution, topic classification, named entity recognition, and summary retrieval. (Son et al., 2022) harnessed annals for developing model that translate Hanja, ancient Korean, into contemporary Korean and English. As for linguistically limited resources, (Lazar et al., 2021) masked language models have been used to predict missing tokens with Oracc Dataset. However, these earlier efforts, predating the recent breakthroughs in LLMs, relied heavily on the accessibility of ancient language training corpora(Assael et al., 2019; Fetaya et al., 2020; Papavassiliou et al., 2020). This resulted in limited resources, constraining the models to a narrow range of specific applications unsuitable for LLM pre-training(Sommerschild et al., 2023). In contrast, we introduce the first ancient Chinese LLM designed to tackle a broad spectrum of downstream tasks, leveraging the extensive and diverse corpus from the PhiloCorpus-ZH and PhiloBenchmark.

3 PhiloGPT

We detail the construction of *PhiloGPT*. Section 3.1 describing the acquisition and organization of *PhiloCorpus-ZH*, which serves as foundation for training. Section 3.2 discuss the development of *PhiloBenchmark*, which comprises various downstream tasks formulated to test the efficacy of LLMs in handling ancient Chinese. In Section 3.3, we explore the *PhiloCoP* framework, which is designed to enhance LLM’s ability to mimic philologists interpreting ancient texts. Section 3.4 provides details with the implementation of *PhiloGPT*.

3.1 PhiloCorpus-ZH

The absence of organized, large-scale ancient corpora has impeded the pre-training of LLMs for ancient languages. Classical Chinese manuscripts, richer and more diverse than inscriptions, provide various genres and content. To bridge this gap, we have curated the extensive *PhiloCorpus-ZH*, systematically categorized from original documents via the Chinese bibliographic *Four-Part Classification* method. Distinctively, our corpus includes previously overlooked folk and semi-folk documents, literature, and socioeconomic records such as legal cases and contracts. These texts, rich in colloquial expressions, capture the linguistic nuances and social dynamics of their times.

Our corpus was sourced from a broad range of publicly available data, including original museum collections, research papers, academic publications, and specialized literature. Distinguished from some previous datasets containing large-scale self-gathered and cleaned web data, our data was derived from the past 40 years (1970s to 2010s) of philology publications that have been meticulously collected, proofread, and curated by our collaborators over years of academic practice. Data cleaning and deduplication were applied to ensure the quality, which included removing extraneous spaces, line breaks, headers, footers, illustrations, tables, formulas, annotation symbols, references and links. More details are available in appendix. Under the close supervision of these experts, we categorized these corpus into four sections:

Chinese Classics: Focuses on the study of ancient Chinese philosophical thoughts, such as interpreting the implications of Confucian classics. These works hold substantial academic value and have historically served as crucial educational and examination materials in ancient times.

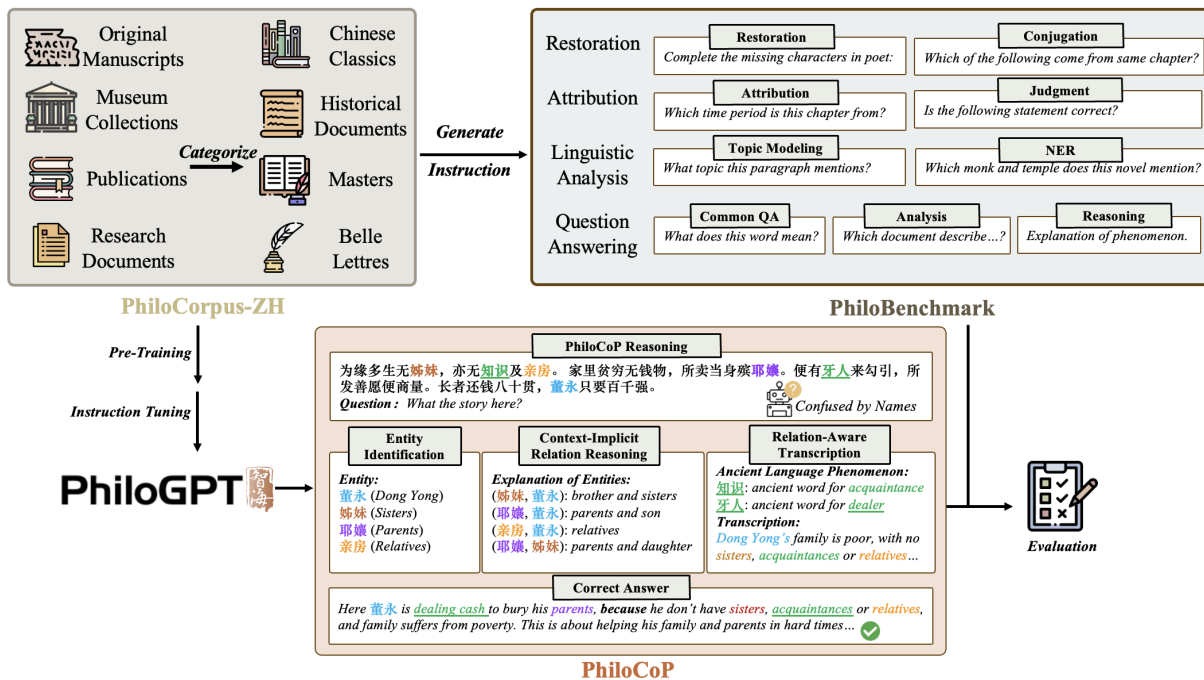


Figure 2: Illustration of PhiloGPT framework.

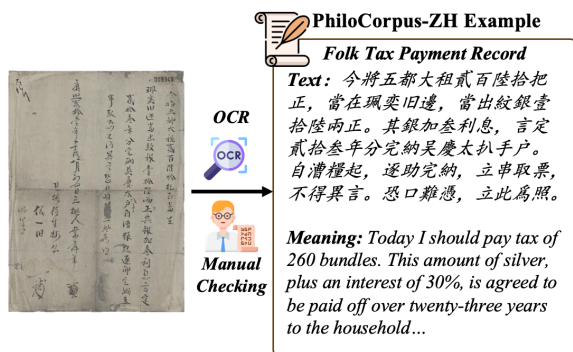


Figure 3: Folk document example of *PhiloCorpus-ZH*.

Historical Documents: Comprises records of historical events and materials, such as official annals in standardized formats, and private records with comments. It also includes other official documents spanning biography, geography, economy and political correspondence.

Masters: Contains notable works and masterpieces that do not easily fit into the other categories. The subjects are diverse, encompassing medicine, astrology, mathematics, arts, and religion.

Belles-lettres: Features a wide range of ancient literary works, such as scriptures, poetry, stories, lyrics and operas, which are popular and widely circulated through manuscripts in ancient times.

Each category encapsulates a broad spectrum of content, topics, and genres, ensuring that our *PhiloCorpus-ZH* exceeds previous corpora in scale,

content richness, and historical breadth. The inclusion of our unique folk corpus expands the existing collection of formal Classical Chinese texts in philology, offering a broader academic scope by addressing issues of word meanings and revealing grammatical contexts more effectively.

3.2 PhiloBenchmark

Philological research demands a nuanced understanding of ancient texts, reflecting the complexity and diversity inherent in these languages. To establish a comprehensive benchmark that enhances and assesses the performance of ancient Chinese LLMs, we have integrated philology methodologies under expert guidance to propose 9 diverse downstream tasks, constituting the **PhiloBenchmark**. *PhiloBenchmark* enables a thorough assessment of an LLM’s ability to understand, interpret, and generate insights from classical Chinese texts. To be specific, our benchmark addresses key areas of mainstream philology studies:

Restoration. Assesses the LLM’s capacity to accurately regenerate missing or damaged manuscript texts, crucial for supporting philologists in document interpretation and evidence alignment. This include directly predicting missing characters (Restoration) and determining whether two passages have a contextual relationship (Conjugation), which is essential for assisting philologists with manuscript recovery through context.

Attribution. Involves determining the origins and historical context of documents, critical for understanding their provenance. Since documentary research, such as Dunhuang manuscript discovery, is concerned with the period in which the writing was transcribed, we introduce the Attribution task for predicting the most likely time period in which this text will appear. We also introduce more sophisticated Judgment tasks for a comprehensive assessment of the model’s ability to attribute in terms of authorship, historical events, etc.

Linguistic Analysis. Tests traditional linguistic abilities for language models such as named entity recognition and segmentation to extract and analyze contextual information from ancient texts.

Question Answering. Measures the LLM’s effectiveness in responding to complex queries tailored to philological research. These tasks can help determine whether model has the potential of facilitating new academic insights.

Given to the severe scarcity of instructional data in ancient Chinese, we implemented hybrid strategies for data generation: (1) *Manual Construction.* For tasks demanding high factual accuracy, such as restoration and attribution, datasets are built from annotated texts and original manuscripts. (2) *Annotated Instruction Expansion.* Leveraging expert philologist insights, we create finely annotated instructions to expand and enhance quality through Self-Instruct(Wang et al., 2022). (3) *Self-QA Enrichment.* We utilize specialized publications and textbooks on ancient Chinese to broaden the scope of instruction data in a Self-QA manner(Bi et al., 2023; Zhang and Yang, 2023a).

Task	Number	Avg. Length
Restoration	50	52
Conjugation	437	352
Attribution	117	76
Judgment	227	88
Topic Modeling	266	106
NER	100	289
Common QA	348	87
Analysis	400	225
Reasoning	500	256

Table 1: Statistics of PhiloBenchmark.

Inspired by the (Wang et al., 2023b) initiative, we employ GPT-4o(Achiam et al., 2023) as a scoring agent to filter out incorrect, low-quality, and irrelevant data. The double-checking for domain

instructions was conducted by philologists and a dedicated team of volunteer students.

3.3 PhiloCoT

Classical Chinese, as a logographic writing system, a single Chinese character can encapsulate multiple meanings that are heavily influenced by context and syntax. Moreover, the limited variety of characters in Classical Chinese gives rise to distinctive grammatical phenomena that are absent in modern Chinese, such as phonetic loans, polysemy, syntactic inversions, and semantic shifts. Furthermore, religious and cultural factors contribute to significant divergences in naming conventions between ancient and contemporary texts. To effectively tackle these complexities, philologists often employ comprehensive interpretation strategies before engaging in downstream tasks.

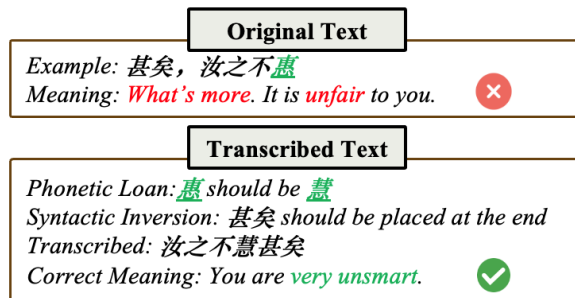


Figure 4: Illustration of typical phenomenon differences in classical Chinese.

Recognizing the profound linguistic disparities between classical and modern Chinese, we developed the *PhiloCoP* (Chain-of-Philology) framework - a multi-step reasoning approach that emulates the analytical processes of philologists when interpreting ancient manuscripts. The essence of PhiloCoP lies in three key stages:

Step 1: Entity Identification. PhiloCoP leverages the information extraction capabilities of LLMs to identify and extract named entities at the character-level. By extracting and categorizing entities, recognizing their significance within the textual structure, the framework establishes a solid foundation for understanding the core components of the ancient manuscript.

Step 2: Context-Implicit Relation Reasoning. Building upon the identified entities, PhiloCoP delves deeper into the connections between the identified entities by uncovering the latent semantic relationships that may not be explicitly stated. The framework constructs a comprehensive network of

relationships, akin to a cognitive map.

Step 3: Relation-Aware Transcription. This step involves generating a standardized transcription of the ancient text that accounts for the unique linguistic characteristics of Classical Chinese. The framework leverages the insights gained from the previous stages to adapt to the semantic shifts, polysemous characters, and grammatical variations prevalent in ancient manuscripts.

By considering the contextual relationships between entities and the nuances of classical language, PhiloCoP produces a reasoning process that faithfully represents the intended meaning of text while mitigating the challenges posed by linguistic anomalies and historical language evolution.

3.4 PhiloGPT

Our *PhiloGPT* model was developed based on Qwen-1.5-7b(Bai et al., 2023) that possesses competitive capabilities in modern Chinese. We conducted pre-training on six Nvidia A800 GPUs, utilizing our domain-specific *PhiloCorpus-ZH*, supplemented by additional open-source general training corpora to prevent catastrophic forgetting. We incorporated a small proportion of domain-specific instructional data to optimize training outcomes in the pre-training stage according to (Zhang and Yang, 2023b). Further training details can be found in the Appendix. For Supervised Fine-tuning, we employed the LoRA tuning method(Hu et al., 2021), focusing on instruction-based supervision to refine the model’s response quality. Our instruction data where constructed following the same strategies as mentioned in Section 3.2, focusing more on generation instead of answering questions in restricted forms (e.g., multiple choices or judgments questions). The ratio of domain-specific to general data was set at 1:5 during pre-training and adjusted to 1:1 during instruction tuning phases. All training processes were executed using the open-source training framework, LLaMA-Factory(Zheng et al., 2024). Comparative experiments were carried out against other generalized pre-trained open-source models of the same parameter scale, providing a fair and meaningful basis for comparison.

4 Experiment

4.1 Evaluation Metric

We assess the performance of LLMs on various philological tasks using the PhiloBenchmark. We specifically compare the performance of our

PhiloGPT with Qwen-7b-chat(Bai et al., 2023), Baichuan2-7b(Yang et al., 2023) and LLaMA2-Chinese-7b(Cui et al., 2023b; Touvron et al., 2023), ensuring models are of the same parameter level. Our evaluation approach is structured into three distinct categories, each aligned with different types of philological tasks: **Restoration.** We employ Character Error Rate (CER) as evaluation metrics in align with (Assael et al., 2022). As for Chronological Attribution, its almost impossible to trace back the exact year, thus we divide time periods according to Chinese dynasty year number and calculation average dynasty shift with ground truth. **Discrimination.** For tasks that involve discrimination and classification, we apply the F1 score and Accuracy as our criterion. This quantifies LLMs’ ability to follow instructions and understand ancient texts. **Generation.** For these generative tasks, we utilize GPT-4o(Achiam et al., 2023) as a scoring agent to perform automatic evaluations. The model assesses the quality of generated content based on factual correctness, accuracy, and richness. From these assessments, we calculate a win rate between competing models, providing a quantitative measure of generative excellence. We take the average results from 3 runs for each task evaluation.

4.2 Experiment Results

Pre-trained on the Dunhuang corpora, *PhiloGPT* exhibits substantial performance improvements across various philology tasks. In Table 2, we benchmark our model against other competitive LLMs. In our study, the ‘-’ in the results for *Restoration*, *Attribution*, and *Conjugation* tasks highlights the deficiencies of open-source LLMs in handling specialized philological tasks. These baseline models, lack of targeted pre-training and fine-tuning specific to philological needs, often generate irrelevant or incorrect responses, or even fail to provide any answer at all (these models may reply "Sorry I don’t know"). The inherent complexities of philological tasks and semantic differences in ancient Chinese, combined with the insufficient training of baseline models on relevant corpora, contribute to their inability to produce viable outcomes in certain cases. For other tasks that involve semi-Classical Chinese, foundation models show some zero-shot capabilities due to the less specialized nature of the language used in these tasks. While improvements in some less challenging tasks are modest, *PhiloGPT* demonstrates

Task	Restoration	Attribution	Conjugation	Topic-M	QA	Judgment
Evaluation Metric	CER↓	Dynasty Shift↓	F1 Score↑	Accu.↑	Accu.↑	Accu.↑
Qwen-7b	-	-	-	29.3%	53.5%	74.2%
Baichuan2	-	-	0.177	23.3%	48.0%	68.9%
LLaMA	-	-	-	10.5%	26.4%	72.3%
PhiloGPT	0.630	1.376	0.451	74.8%	62.1%	77.5%
PhiloGPT+CoP	0.579	1.305	0.590	75.6%	65.2%	86.7%

Table 2: Results of LLM performance with PhiloBenchmark evaluation. The '-' denotes that LLM could not solve the philology specific task, i.e., refuse to answer or randomly generate answers. For Restoration and Attribution, the smaller score means better performances, denoted as a ↓ besides evaluation metric. While for other tasks, greater value indicates better, denoted as a ↑. The best result for PhiloGPT+CoP is bold.

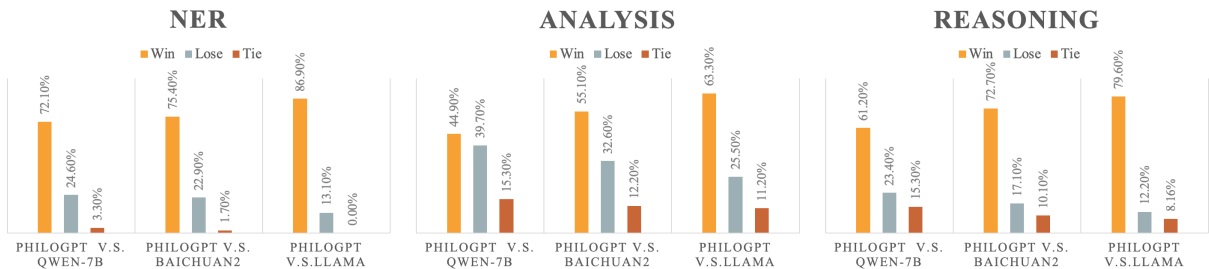


Figure 5: The Evaluation Results of Reasoning and Analysis Tasks, which uses GPT-4o judged win rate as evaluation metric, the greater win rate showcases better capability.

notable enhancements in tasks centered around analytical QA, generation and retrieval.

Implementing *PhiloCoP* has led *PhiloGPT* to performance gains across various tasks. *PhiloCoP* transcribe the original documents to eliminate lexical biases and confusions inherent in Classical Chinese grammar, and extracted key information through NER to explicitly. This method has yielded performance enhancements in *PhiloGPT*. For instance, when tackling analytical tasks, philologists deduce and hypothesize based on keyword usage, names of typical people and places, and their relationships. *PhiloCoP* emulates such practice, resulting in improvements in these tasks. However, applying *PhiloCoP* to LLMs not pre-trained on ancient texts yields counterproductive effects. This is because generalized LLMs lack the capacity to capture latent semantic relationships through entity identification and transcription, thereby reverting to unreliable answers based on modern Chinese guesses. This highlights *PhiloGPT*'s genuine understanding of ancient texts and its ability to perform downstream tasks more effectively.

Our *PhiloBenchmark*, developed from diverse topics and sources, provides a thorough assessment of ancient LLMs from a philological perspective. While generalized LLMs demonstrate respectable outcomes for basic capabilities,

PhiloGPT distinctly outperforms in specialized philology tasks, such as restoration, highlighting the tailored advantages of our model.

5 Dunhuang Manuscript Case Study

Taking a step further, we conducted an analysis using Dunhuang manuscripts as a case study to explore two specialized tasks of philology research practice. Our objective is to show that, following pre-training and instruction tuning, *PhiloGPT* can provide researchers with enhanced flexibility to perform comprehensive text information mining. This approach not only facilitates deeper textual analysis but also leads to potential discoveries.

5.1 Analyze Copying Relationships

We selected Dunhuang manuscripts for our case study because they are among the most renowned, numerous, and cover the broadest span of time. Within this vast collection, several manuscripts demonstrate instances of duplication and copying relationships. Establishing the chronological order of these texts is a crucial task, as the original version can reflect the more realistic original appearance of the manuscript. As illustrated in Figure 6 (a), we employed *PhiloGPT* to analyze their interrelationships by context reasoning, which is the **homoeoteleuton** phenomenon. This knowledge

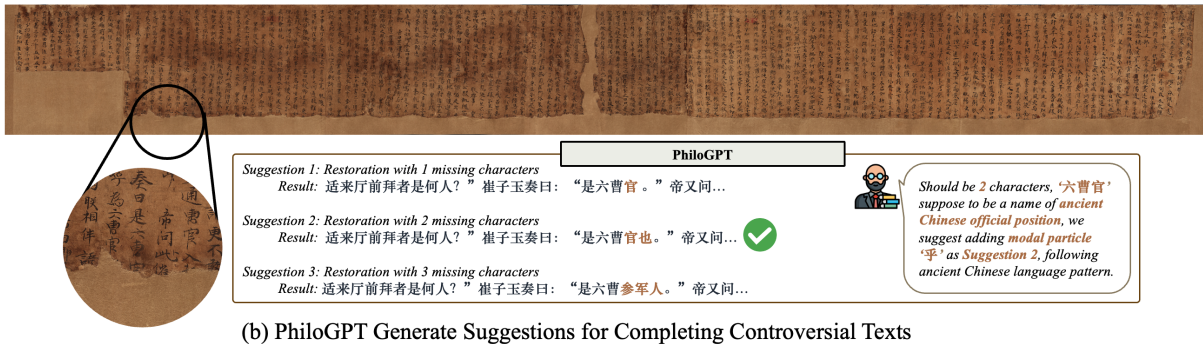
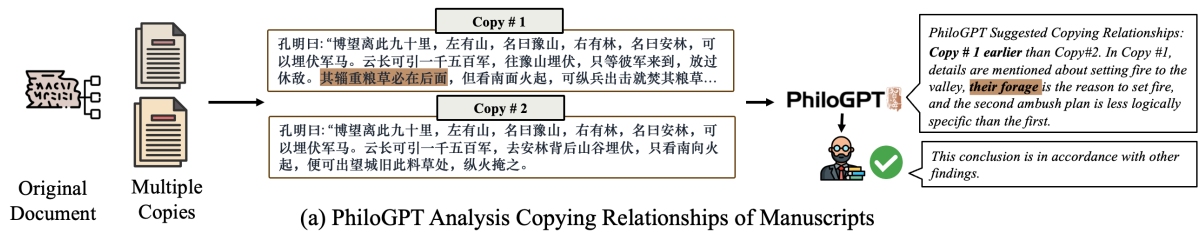


Figure 6: Illustration of PhiloGPT for text discoveries with Dunhuang as case study.

is not in the foundation model large-scale storage, but learned and generalized to the new applications since Qwen-7b base model fails in such cases.

5.2 Generate Suggestion for Text Criticism

The ancient Chinese novel represents a unique genre, showcasing a level of innovation and imagination surpassing its contemporaries. These works were often viewed as unconventional by the rulers of their times, exemplified by classics such as 'Journey to the West.' Restoration of such texts poses significant challenges for philologists, particularly due to the scarcity of reference materials, which are crucial for identifying parallels (recurring expressions and linguistic peculiarities) for text restoration. As illustrated in Figure 6 (b), this manuscript employs a narrative to satirize the emperor of the era. Unfortunately, the original manuscript is damaged at the bottom, leading to disagreements regarding the extent of the missing text.

To address this, we utilized *PhiloGPT* to generate several plausible completions for the missing sections. The restoration proposal consisting of three missing characters was deemed most credible (in line with the overall style and language conventions of the article). While these results still require validation from philology, our model has demonstrated potential as a valuable tool for assisting philologists in making new scholarly discoveries.

6 Conclusion

In this paper, we introduce *PhiloGPT*, the first domain-specific LLM designed to tackle a range of tasks related to the discovery of ancient Chinese manuscripts. We developed the large-scale *PhiloCorpus-ZH*, which encompasses diverse topics of original ancient texts and associated scholarly research. To align with the unique characteristics of these languages, we proposed the *Philo-CoP* framework. From a philology perspective, our *PhiloBenchmark* serves as an evaluation standard for LLMs tailored to ancient Chinese. Extensive experiments and case studies with Dunhuang manuscripts have demonstrated *PhiloGPT*'s potential for ancient document discovery. For future work, we may extend *PhiloGPT*'s capability based on more applications and feedback when applied in other manuscripts. We also consider incorporating Retrieval-Augmented Generation (RAG) with specific documents to mitigate hallucination problems.

Limitation: Due to historical reasons, official documents are often better preserved and more complex than folk texts, resulting in some data bias and imbalance. We anticipate that the discovery of additional manuscripts, such as those recently unearthed in Xinjiang, will help mitigate this issue. Moreover, ancient documentary studies place a high premium on factual accuracy. Thus, findings from our model still require secondary verification by philologists and must undergo peer review. Additionally, considering that manuscripts often incorporate image modality, fine-tuning a multimodal

LLM that integrates text styles with document materials could enhance accuracy in attribution tasks. Multi-modality input would provide scholars with greater flexibility when deploying LLMs.

Acknowledgements

This work was in part supported by Special Funds of the National Natural Science Foundation of China (No. 62441605), National Science and Technology Major Project (2022ZD0119100), National Natural Science Foundation of China (No. 72342023, 62402429), National Social Science Fund of China (No. 21BYY142, 14AZS001), Key Research and Development Program of Zhejiang Province (No. 2024C03270), the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010). The authors express their deep gratitude to Zheqi Lv for his invaluable advice and to Karl Cao for his technical support.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Yannis Assael, Thea Sommerschild, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on greek epigraphy. *arXiv preprint arXiv:1910.06262*.
- Yannis Assael, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2023. OceanGPT: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.
- Jiayi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.
- Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.
- Imre Galambos. 2020. *Dunhuang Manuscript Culture: End of the First Millennium*, volume 22. Walter de Gruyter GmbH & Co KG.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.
- Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, Sujin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the joseon dynasty via neural language modeling and machine translation. *arXiv preprint arXiv:2104.05964*.
- Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in ancient akkadian texts: a masked language modelling approach. *arXiv preprint arXiv:2109.04513*.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Katerina Papavassileiou, Dimitrios I Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in infilling text from ancient tablets. *ACM Journal on Computing and Cultural Heritage*, 16(3):1–25.

- Katerina Papavassiliou, Gareth Owens, and Dimitrios Kosmopoulos. 2020. A dataset of mycenaean linear b sequences. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2552–2561.
- Clifford Seth Parker, Stephen Parsons, Jack Bandy, Christy Chapman, Frederik Coppens, and William Brent Seales. 2019. From invisibility to readability: recovering the ink of herculaneum. *PLoS one*, 14(5):e0215775.
- Thea Sommerschildt, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.
- Juhee Son, Jiho Jin, Haneul Yoo, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Translating hanja historical documents to contemporary korean and english. *arXiv preprint arXiv:2205.10019*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Hutu: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.
- Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023b. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhakaran Kambar, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.
- Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Ding-gang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.
- Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, et al. 2024. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *arXiv preprint arXiv:2404.02893*.
- Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Haneul Yoo, Jiho Jin, Juhee Son, JinYeong Bak, Kyunghyun Cho, and Alice Oh. 2022. Hue: Pretrained model and dataset for understanding hanja documents of ancient korea. *arXiv preprint arXiv:2210.05112*.
- Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.
- Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382.
- Xuanyu Zhang and Qing Yang. 2023a. Self-qa: Unsupervised knowledge guided language model alignment. *arXiv preprint arXiv:2305.11952*.
- Xuanyu Zhang and Qing Yang. 2023b. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4435–4439.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Hyper-parameter	Setting
Fine-tuning Type	Full & LoRA
Cutoff Length	2048
Learning Rate	1e-4
Training Epoch	Full: 1, LoRA:8
Batch Size	32
Optimizer	adamw_torch
LoRA γ	8
LoRA α	16
LoRA Dropout	0.1
LoRA Target	All

Table 3: Hyper-parameters of training PhiloGPT.

A Training Details

In this section, we give an example of the Hyper-parameters of our PhiloGPT training details.

For generalized pre-training dataset, we sampled from Wikipedia (en), Wikipedia (zh), SkyPile (zh), Wanjuan 1.0. For hybriding generalized Supervised Fine-Tuning dataset, we sampled from Stanford Alpaca (zh), BELLE 2M (zh), BELLE School Math 0.25M, Alpaca CoT (multilingual), Firefly 1.1M (zh), Web QA (zh), ShareGPT4 (en&zh), Ruozhiba (zh).

B Experiment Costs

We conducted experiments with GPT4 API, utilizing the gpt-4-turbo-2024-04-09 model, at a cost of \$5 per 1 million tokens input and \$15 per 1 million output. We spent \$240 in total. We conducted data cleaning with the guidance of philology experts in our co-authors, and with a group of student volunteers from the college of Computer Science and the college of liberal arts.

C PhiloCoP Prompt Template

Here presents our PhiloCoP template for PhiloGPT when performing our multi-step reasoning. For reading convenience, we have translated our prompt to English. In practice, we use the original Chinese prompt, being consistent with all LLM proposed output language.

D Prompting Template for GPT-4o

Here presents our template for GPT-4o usage as a scoring agent for evaluating generation tasks. For reading convenience, we have translated our prompt to English. In evaluation practice, we use the original Chinese prompt, being consistent with

PhiloCoP Prompt Template

Step 1: Entity Identification.

Find all entities based on provided text.

{ *Original Ancient Text.* }

Step 2: Context-Implicit Relation Reasoning.

Explain relation given a set of entities and text.

{ *Original Ancient Text.* }

{ *Entities.* }

Step 3: Relation-Aware Transcription.

Transcribe all text given entities and relations, then answer my question.

{ *Original Ancient Text.* }

{ *Entities with Relationship.* }

{ *Question.* }

Table 4: Prompt for evaluating generation tasks.

our PhiloGPT proposed language. Additionally, we randomly switch the answer of PhiloGPT from position A to B to avoid bias.

Prompt for Evaluating Generation Tasks

Here is a question { *question* }.

Here is a sample answer { *sample answer* }.

Answer from A: { *output from first LLM* }.

Answer from B: { *output from second LLM* }.

Please decide which of these two answers is better. Reply 'A' or 'B' only. Try your best for making decisions. If you feel very unsure, just reply 'None'.

Table 5: Prompt for evaluating generation tasks.

E Example of PhiloCorpus-ZH

Here we provide more details and samples of our PhiloCorpus-ZH data in Table 7-9. We give each category a sample with the original ancient Chinese text.

F Example of PhiloBenchmark

Here we provide more details and samples of our PhiloBenchmark data in Table 10-11. We give each category a sample with the original ancient Chinese text. For reading convenience, we have translated the instructions and output to English if available.

G PhiloGPT on Open-ended Questions with Dunhuang

To fully illustrate PhiloGPT capabilities in answering open-ended questions with real-world application, we further give more showcases in Table 12. These questions are open-ended and require complex reasoning and understanding of ancient text, which were used as Final Examination Questions for college of liberal arts students.

PhiloCorpus-ZH Samples Categorized by Topics

Chinese Classics

Confucius Classics

子曰：“学而时习之，不亦说乎？有朋自远方来，不亦乐乎？人不知而不愠，不亦君子乎？”有子曰：“其为人也孝弟，而好犯上者，鲜矣；不好犯上而好作乱者，未之有也。君子务本...

Poetry Classics

关关雎鸠，在河之洲。窈窕淑女，君子好逑。参差荇菜，左右流之。窈窕淑女，寤寐求之。求之不得，寤寐思服。悠哉悠哉，辗转反侧。参差荇菜，左右采之。窈窕淑女，琴瑟友之。

Characters

天地玄黄，宇宙洪荒。日月盈昃，辰宿列张。寒来暑往，秋收冬藏。闰余成岁，律吕调阳。云腾致雨，露结为霜。金生丽水，玉出昆冈。剑号巨阙，珠称夜光。果珍李柰，菜重芥姜...

Phonology

帝高阳之苗裔兮，朕皇考曰伯庸。摄提贞于孟陬兮，惟庚寅吾以降。皇览揆余初度兮，肇锡余以嘉名。名余曰正则兮，字余曰灵均。纷吾既有此内美兮，又重之以修能。扈江离与辟芷兮...

Ritual

大学之道，在明明德，在亲民，在止于至善。知止而后有定，定而后能静，静而后能安，安而后能虑，虑而后能得。物有本末，事有终始。知所先后，则近道矣。古之欲明明德于天下者，先治其国。欲治其国者，先齐其家。欲齐其家者，先修其身。欲修其身者，先正其心...

Virtue

仲尼居，曾子侍。子曰：“先王有至德要道，以顺天下，民用和睦，上下无怨。汝知之乎？”曾子避席曰：“参不敏，何足以知之？”子曰：“夫孝，德之本也，教之所由生也。复坐，吾语汝。身体发肤，受之父母，不敢毁伤，孝之始也。立身行道，扬名于后世，以显父母...

Historical Documents

Official History

项羽既定河北，率诸侯兵欲西入关。先是，诸侯吏卒、繇使、屯戍过秦中者，秦中吏卒遇之多无状。及章邯以秦军降诸侯，诸侯吏卒乘胜多俘虏使之，轻折辱秦吏卒。秦吏卒多怨，窃言曰：“章将军等诈吾属降诸侯。今能入关破秦，大善；即不能，诸侯虏吾属而东，秦又尽诛吾父母妻子...

Annal

起著雍摄提格正月，尽七月，不满一年。高祖神尧大圣光孝皇帝上之上。武德元年，戊寅公元 618 年。春，正月，丁未朔，隋恭帝诏唐王剑履上殿，赞拜不名。唐王既克长安，以书谕诸郡县，于是东自商洛，南尽巴、蜀，郡县长吏及盗贼渠帅、氏羌酋长，争遣子弟入见请降，有司复书...

Miscellaneous

张仪欲假秦兵以救魏。左成谓甘茂曰：“子不予之。魏不反秦兵，张子不反秦。魏若反秦兵，张子得志于魏，不敢反于秦矣。张子不去秦，张子必高子。”

Biography

庄公矜勇力不顾行义晏子谏。庄公奋乎勇力，不顾于行义。勇力之士，无忌于国，贵戚不荐善，逼逐不引过，故晏子见公。公曰：“古者亦有徒以勇力立于世者乎？”晏子对曰：“婴闻之，轻死以行礼谓之勇，诛暴不避谓之力。故勇力之立也，以行其礼义也。汤武用兵而不为逆...

Geography

... 有慈恩道场三藏法师，讳玄奘，俗姓陈氏，其先颍川人也。帝轩提象，控华渚而开源；大舜宾门，基历山而耸构。三恪照于姬载，六奇光于汉祀。书奏而承朗月，游道而聚德星...

Table 6: PhiloCorpus-ZH Samples (Continuous).

PhiloCorpus-ZH Samples Categorized by Topics

Government Officials

吏部尚书一人，正三品；（周之天官卿也。《汉旧仪》云：“尚书四人，为四曹：一曰常侍曹，二曰二千石曹，三曰民曹，四曰客曹。成帝增置三公曹，为五曹。其常侍曹主丞相、御史、公卿事。后汉光武又分为六曹；常侍曹为吏部曹，主选举、斋祀事。”）

Administrative Record

... 知此三者，谓之治政。夫地载而不弃也，一著而不迁也，安固而不动，则莫不生殖。圣人因之设井邑，列比闾，使察黎民之数，赋役之制，昭然可见也。自秦孝公用商鞅计...

Commentary

自有王者，便置诸侯，列以五等，疏为万国。当周之东迁，王室大坏，于是礼乐征伐自诸侯出。迄乎秦世，分为七雄。司马迁之记诸国也，其编次之体，与本纪不殊。盖欲抑彼诸侯，异乎天子，故假以他称，名为世家。

Categorization

（经学、小学书，以国朝人为极，于前代著作，撷长弃短，皆已包括其中，故于宋元明人从略。）正经正注第一（此为诵读定本，程试功令，说经根柢。注疏本与明监本五经，功令并重。）《十三经注疏》。（共四百一十六卷。乾隆四年武英殿刻附考证本，同治十年广州书局覆刻殿本...

Masters

Canonical Works

人之初，性本善。性相近，习相远。苟不教，性乃迁。教之道，贵以专。昔孟母，择邻处。子不学，断机杼。窦燕山，有义方。教五子，名俱扬。养不教，父之过。教不严，师之惰...

Military Strategy

孙子曰：兵者，国之大事，死生之地，存亡之道，不可不察也。故经之以五事，校之以计，而索其情：一曰道，二曰天，三曰地，四曰将，五曰法。道者，令民与上同意也，故可与之死，可与之生，而不畏危也。天者，阴阳、寒暑、时制也。地者，远近、险易、广狭、死生也...

Legalism

世之显学，儒、墨也。儒之所至，孔丘也。墨之所至，墨翟也。自孔子之死也，有子张之儒，有子思之儒，有颜氏之儒，有孟氏之儒，有漆雕氏之儒，有仲良氏之儒，有孙氏之儒，有乐正氏之儒...

Agriculture

凡稻种最多。不粘者，米曰粳。粘者，米曰糯。（南方无粘黍，酒皆糯米所为。）质本粳而晚收带粘（俗名婺源光之类）不可为酒，只可为粥者，又一种性也。凡稻谷形有长芒、短芒、（江南名长芒者曰浏阳早，短芒者曰吉安早。）长粒、尖粒、圆顶、扁圆面不一。

Medicine

人参，亦名黄参、血参、人衔、鬼盖、神草、土精、地精、海腴、皱面还丹。气味（根）甘、微寒、无毒。主治阴亏阳绝之症（昏眩、自汗、痰鸣、脉大）。用人参十两细切，加水二十碗浸透，以桑柴火缓煎成膏。每服一至三碗。持续服至病愈。以方名“人参膏”。胸中痞坚，胁下逆气抢心...

Mathematics

凡大数之法：万万曰亿，万万亿曰兆，万万兆曰京，万万京曰垓，万万垓曰秭，万万秭曰沟，万万沟曰涧，万万涧曰正，万万正曰载。

Astrology

老子曰：无名天地之始，有名万物之母，有物混成先天地生列。列御寇曰：有形生於无形，天地之初，有太易，有太初，有太始，有太素，太易者，气未见。太初者，气之始。太始者，形之始。太素者，质之始。气与形质，合而未离。

PhiloCorpus-ZH Samples Categorized by Topics

Arts

余尝以隐显、回互、激射说古文，然行文之法，又有奇偶、疾徐、垫拽、繁复、顺逆、集散。不明此六者，则于古人之文，无以测其意之所至。而第其诣之所极，垫拽繁复者，回互之事...

Ceremony

茶者，南方之嘉木也，一尺二尺，乃至数十尺。其巴山峡川有两人合抱者，伐而掇之，其树如瓜芦，叶如梔子，花如白蔷薇，实如^①桐，蒂如丁香，根如胡桃。

Naming

「张」源出；一「姬」姓，据《通志·氏族略·以字为氏》所载，张氏，世代仕晋，晋分为三，又世代仕韩，后渐成望族。二、改姓，三国魏张辽原姓「聂」，改姓「张」。又三国蜀诸葛亮赐南蛮酋长龙佑那为「张」氏。三、「张」氏自汉代以后，人数渐多，与当时道教的兴起与流行有关...

Encyclopedia

《三五历记》曰：未有天地之时，混沌状如鸡子，溟幸始牙，岁在摄提，元气肇始。又曰：清轻者上为天，浊重者下为地，冲和气者为人。故天地含精，万物化生。

Novel

陈仲举言为士则，行为世范，登车揽辔，有澄清天下之志。为豫章太守，至便问徐孺子所在，欲先看之。主簿曰：“群情欲府君先入廨。”陈曰：“武王式商容之闾席不暇暖。吾之礼贤，有何不可！”

Belle-lettres

Buddhist Scripture

佛言：“善哉，善哉！须菩提！如汝所说：‘如来善护念诸菩萨，善付嘱诸菩萨。’汝今谛听，当为汝说。善男子、善女人，发阿耨多罗三藐三菩提心，应如是住，如是降伏其心。”

Poetry

君不见建章宫中金明枝，万万长条拂地垂。二月三月花如霰，九重幽深君不见。艳彩朝含四宝宫，香风吹入朝云殿。汉家宫女春未阑，爱此芳香朝暮看。看看看来心不忘，攀折将安镜台上...

Stories

明妃既策立，元来不称本情；可汗将为情和，每有善言相向。“异方歌乐，不解奴愁；别域之欢，不令人爱。”单于见他乐，又传一箭，告报诸蕃，非时出猎，围绕烟口山，用昭君作心，万里攒军，千兵逐兽。昭君既登高岭，愁思便生，遂指天叹帝乡而曰处，若为陈说...

Lyrics

古英雄未遇时，都无大志，非止邓禹希文学、马武望督邮也。晋文公有妻有马，不肯去齐。光武贫时，与李通讼逋租于严尤。尤奇而目之。光武归谓李通曰：“严公宁目君耶”。窥其意...

Opera

古董先生谁似我？非玉非铜，满面包浆裹。剩魄残魂无伴侣，时人指笑何须躲。旧恨填胸一笔抹，遇酒逢歌，随处留皆可。子孝臣忠万事妥，休思更吃人参果。

Seperate Collections

古之有天下者，以为大不得已，而后世以为乐。此天下所以难有也。生民之初，固无乐乎，为君不幸，为天下所归，而不可得拒者，天下有求于我，我无求于天下也。子不闻至德之世乎，饭糗梁啜藜藿，饮食未侈也；夏葛衣冬鹿裘，衣服未备也；土阶三尺茅茨不穷，宫室未美也。

Table 8: PhiloCorpus-ZH Samples (Continuous).

PhiloBenchmark Samples

Restoration

Instruction: 请填写空缺位置所代表的字 (Fill in the blank works in given paragraph):

Input: 人生在世审思量，暂时吵闹有何方？大众志心须净听，先须孝顺阿耶。好事恶事皆抄录，善恶童子每抄将。孝感先贤说董永，年登十五二亲亡。自叹福薄无兄弟，____ 数千行。

Output: 人生在世审思量，暂时吵闹有何方？大众志心须净听，先须孝顺阿耶。好事恶事皆抄录，善恶童子每抄将。孝感先贤说董永，年登十五二亲亡。自叹福薄无兄弟，眼中流泪数千行。

Conjugation

Instruction: 请判断下面两段内容是否出自同一写卷/文章 (Decide whether these two paragraphs are from the same manuscript/document)

Input: 第一段 (Paragraph 1): 明妃既策立，元来不称本情；可汗将为情和，每有善言相向。“异方歌乐，不解奴愁；别域之欢，不令人爱。”单于见他 不乐，又传一箭，告报诸蕃，非时出猎，围绕烟山，用昭君作心，万里攒军，千兵逐兽。

第二段 (Paragraph 2): 昭君一度登山，千回下泪，慈母只今何在？君王不见追来。当嫁单于，谁望喜乐？良由画匠，捉妾陵持，遂使望断黄沙，悲连紫塞，长辞赤县，永别神州。

Output: 出自同一卷 (Same paragraph)

Attribution

Instruction: 请判断下面这段文字的写作年代 (Identify which time period this paragraph from):

Input: “药王！若有菩萨闻是法华经，惊疑、怖畏，当知是为新发意菩萨；若声闻人闻是经，惊疑、怖畏，当知是为增上慢者。“药王！若有善男子、善女人，如来灭后，欲为四众说是法华经者，云何应说？是善男子、善女人，入如来室，著如来衣，坐如来座，尔乃应为四众广说斯经。如来室者，一切众生中大慈悲心是；如来衣者，柔和忍辱心是；如来座者，一切法空是。安住是中，然后以不懈怠心，为诸菩萨及四众广说是法华经。“药王！我于馀国，遣化人为其集听法众，亦遣化比丘、比丘尼、优婆塞、优婆夷听其说法，是诸化人，闻法信受，随顺不逆。若说法者在空闲处，我时广遣天、龙、鬼神、乾闥婆、阿修罗等，听其说法。我虽在异国，时时令说法者得见我身。若于此经忘失句逗，我还为说，令得具足。”

Output: 吐蕃统治时期写本 (Manuscripts from Tibetan rule time.)

NER

Instruction: 指出下面内容中的所有地点、人名、年代等命名实体 (Identify the named entities in following text, including locations, persons, date and time...):

Input: 《下女夫词》共有十多个抄本，它是用对答对歌的文学手法，详细记录了民间结婚程式的过程：女夫先到女家大门，女方请新郎下马，在男方向女方奉献了绶罗之后，女方向新郎敬上“上门酒”，并诵《上酒诗》，男方要作答诗。

Output: 《下女夫词》，女夫，绶罗，上门酒，《上酒诗》

Topic Modeling

Instruction: 请指出下面文字内容的主题符合 10 类中哪一类 (社会经济、宗教、文化艺术、...) (Identify the topic of following text from 10 candidates: economics, religion, culture and arts, ...)

Input: 羽葆本为仪仗名，以鸟羽为饰物，隋唐时诸王大臣有功者往往加羽。唐乐中羽葆为鼓吹的一部，其所奏有十八曲：太和、体和、七德、驺虞、基王化、篆唐风、厌炎精、肇皇运、躡龙飞、殄马邑、兴晋阳、济润险、应圣期、御底极、宁兆庶、服遐荒、龙池、破阵乐，均为歌颂唐王朝功业的乐出。

Output: 文化艺术 (Culture and arts)

Table 9: PhiloBenchmark Samples.

PhiloBenchmark Samples

Judgment

Instruction: 在《文样·贤兄》中，词语“冰镜”被用来比喻楷模和榜样，“冰镜照妍丑”用来描述光滑的冰面如镜子般地能够照见人的容貌，因此可以说“冰镜”在这段文字中的含义具有象征意义，指的是衡量标准或模范。(Identify Whether the description of word in a specific context is correct.)

Input:

Output: 正确 (Correct)

Common QA

Instruction: “无量劫”一词在敦煌文献中如何描述? A: 佛教语，形容极其的时间 B: 佛教语，形容时间有限 C: 佛教语，形容平常的时间 D: 佛教语，形容即刻的时间" (How is the term '无量劫' described in Dunhuang manuscripts? A: Buddhist term, describing an extremely long period of time. B: Buddhist term, describing a limited period of time. C: Buddhist term, describing a common period of time. D: Buddhist term, describing an immediate period of time.)

Input:

Output: B

Analysis

Instruction: 弥勒经变起源于哪个时期，主要反映了什么样的佛教思想? (From which period does the Maitreya Sutra Transformation originate, and what kind of Buddhist thought does it primarily reflect?)

Input:

Output: 弥勒经变起源于唐代，其主要反映了对未来佛弥勒的崇拜和对未来世界的美好的期望。这类壁画通常表现了弥勒佛降生、人间极乐以及未来世界的种种美景，与唐代社会对未来理想国度的向往相契合。在佛教艺术中，弥勒经变不仅是一种宗教表达，更是一种文化寄托，展示了当时人们对来世的憧憬与信仰。(The Maitreya Sutra Transformation originated during the Tang Dynasty and primarily reflects the worship of Maitreya, the future Buddha, and the anticipation of a better future world. These murals typically depict the birth of Maitreya Buddha, the earthly paradise, and various beautiful scenes of the future world, aligning with the Tang society's longing for an ideal future. In Buddhist art, the Maitreya Sutra Transformation is not only a form of religious expression but also a cultural aspiration, showcasing people's dreams and beliefs about the afterlife at that time.)

Reasoning

Instruction: 请帮我解释原文中“都满”的意思。(Explain the word meaning "都满" in this context.)

Input: 都满阁梨道德高，音律清冷能宛转。好韵宫商申雅调，高着声音唱将来。

Output: 魏晋以后，佛家讲经，由一人经、问难，一人释义、答难，前者称“都满”，后者称“法師”。唐代俗讲，负责唱经者也称“都请”。

Table 10: PhiloBenchmark Samples (Continuous).

PhiloGPT Showcases on open-ended questions with Dunhuang case studies

Instruct: 如何鉴别敦煌写本的真伪，同时如何合理看待疑似伪造的敦煌卷子？(How to identify the authenticity of manuscripts and how to reasonably regard suspected forged Dunhuang scrolls?)

PhiloGPT Output: 鉴别敦煌写本的真伪需要从写本内容、题记和收藏印三个方面综合考虑。要仔细检验纸张、书法、印鉴等外观特征，并结合写卷本身的历史和典籍知识进行判断。对于疑似伪造的敦煌卷子，不能轻易否定其价值，需进一步研究其来历和传承经过。在学术研究中，务必确保基础资料的真实性，避免误导。(To identify the authenticity of Dunhuang manuscripts, it is necessary to comprehensively consider the content of the manuscripts, inscriptions, and collection seals. One should carefully examine the physical characteristics such as the paper, calligraphy, and seals, and make judgments based on the historical context and bibliographical knowledge of the manuscript itself. For suspected forged Dunhuang scrolls, one should not easily dismiss their value but rather further investigate their origins and provenance. In academic research, it is essential to ensure the authenticity of fundamental materials to avoid misleading conclusions.)

Instruction: 在敦煌写本的多装帧样式中，列举出让你印象深刻的 5 种样式，并简要论述每一种样式的优点和缺点。(Among the various binding styles of Dunhuang manuscripts, list five styles that impressed you the most, and briefly discuss the advantages and disadvantages of each style.)

PhiloGPT Output: 1. 卷轴装：这是一种最古老的装帧样式，用于保存纸张。它易于使用，可以轻松翻阅，但需要一定的维护和保养，以防止纸张断裂和破损。

2. 经折装：这是一种在纸张折叠后制成的书的装帧样式。它便于携带和保存，但需要一定的保养，以防止折叠处破损。

3. 册子装：册子装通常用于书籍，它将多张纸张粘在一起，形成一个完整的书本。它比卷轴装更耐用，但不太适合翻阅。

4. 梵夹装：这是一种将纸张折叠成小书本的装帧样式。它比册子装更轻便，但不太适合保存。

5. 蝴蝶装：蝴蝶装是一种折叠的书的装帧样式，它将纸张折叠成蝴蝶形状。它非常轻便，但同样容易损坏。

(1. Scroll Binding: This is one of the oldest binding styles used to preserve paper. It is easy to use and can be easily perused, but it requires a certain level of maintenance to prevent the paper from breaking and tearing.

2. Accordion Binding: This style involves folding the paper to create a book. It is convenient to carry and store, but requires some care to prevent damage at the folds.

3. Booklet Binding: Commonly used for books, this style glues multiple sheets of paper together to form a complete book. It is more durable than scroll binding but less suitable for frequent perusal.

4. Fan-fold Binding: This style folds the paper into small booklets. It is more portable than booklet binding but less suitable for long-term preservation.

5. Butterfly Binding: This style folds the paper into a butterfly shape. It is very lightweight but also easily damaged.)

Table 11: PhiloGPT showcases.