AC-EVAL: Evaluating Ancient Chinese Language Understanding in Large Language Models

Yuting Wei[*], **Yuanxing Xu**[*], **Xinru Wei, Simin Yang, Yangfu Zhu,**
Yuqing Li, Di Liu, Bin Wu[†]
Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia,
Beijing University of Posts and Telecommunications
{yuting_wei, xyx, wxr2000, ysm2000, zhuyangfu, liyuqing, liudi, wubin}@bupt.edu.cn

Abstract

Given the importance of ancient Chinese in capturing the essence of rich historical and cultural heritage, the rapid advancements in Large Language Models (LLMs) necessitate benchmarks that can effectively evaluate their understanding of ancient contexts. To meet this need, we present AC-EVAL, an innovative benchmark designed to assess the advanced knowledge and reasoning capabilities of LLMs within the context of ancient Chinese. AC-EVAL is structured across three levels of difficulty reflecting different facets of language comprehension: general historical knowledge, short text understanding, and long text comprehension. The benchmark comprises 13 tasks, spanning historical facts, geography, social customs, art, philosophy, classical poetry and prose, providing a comprehensive assessment framework. Our extensive evaluation of top-performing LLMs, tailored for both English and Chinese, reveals a substantial potential for enhancing ancient text comprehension. By highlighting the strengths and weaknesses of LLMs, AC-EVAL aims to promote their development and application forward in the realms of ancient Chinese language education and scholarly research.[1]

1 Introduction

The advent of Large Language Models (LLMs) has significantly impacted Natural Language Processing (NLP), highlighting their importance in understanding and generating human languages (Wei et al., 2022a; Zhou et al., 2022; Zhao et al., 2023). With the rise of Chinese as a major global language, there has been a surge in Chinese-specific LLMs (Zeng et al., 2022; Bai et al., 2023; Baichuan, 2023). Ancient Chinese, a crucial part of the Chinese language, records a rich historical and cultural heritage, and has garnered considerable attention from computational linguists (Li et al., 2022; Wang et al., 2023). LLMs present significant opportunities for enhancing the pedagogy of Chinese literary education through convenient text analysis and comprehension. Therefore, assessing the ancient Chinese comprehension capabilities of LLMs holds significant importance.

Initially, benchmarks for LLMs primarily targeted the assessment of English language understanding, exemplified by MMLU (Hendrycks et al., 2021), BIG-bench (Srivastava et al., 2023) and HELM (Liang et al.). Subsequently, several benchmarks focusing on Chinese, such as C-Eval (Huang et al., 2023), CMMLU (Li et al., 2024), and SuperCLUE (Xu et al., 2023), were introduced. These benchmarks aim to evaluate the reasoning performance of LLMs across a broad spectrum of fields including STEM, social sciences, and humanities. However, these benchmarks tend to lean towards modern Chinese comprehension. While some include tasks related to Chinese language, literature and history, they are often relegated to minor categories, insufficient for a comprehensive coverage of ancient Chinese knowledge and language assessment. Existing benchmarks for ancient Chinese understanding, such as CCLUE [2] and WYWEB (Zhou et al., 2023), cover various aspects but primarily focus on linguistic feature analysis, frequently overlooking the assessment of historical knowledge hidden in literature. Furthermore, the diversity in format across these datasets, tailored for specific tasks rather than providing a unified assessment framework, complicates the evaluation of LLMs, presenting challenges in conducting uniform assessments.

To bridge this gap, we propose AC-EVAL (as illustrated in Figure 1), a benchmark meticulously designed for a comprehensive evaluation of LLMs'

---

[*] Equal contribution.

[†] Corresponding author.

[1] The AC-EVAL data and evaluation code are available at `https://github.com/yuting-wei/AC-EVAL`.

[2] https://github.com/Ethan-yt/guwen-models

Findings of the Association for Computational Linguistics: EACL 2024, pages 1600–1617
November 12-16, 2024 ©2024 Association for Computational Linguistics

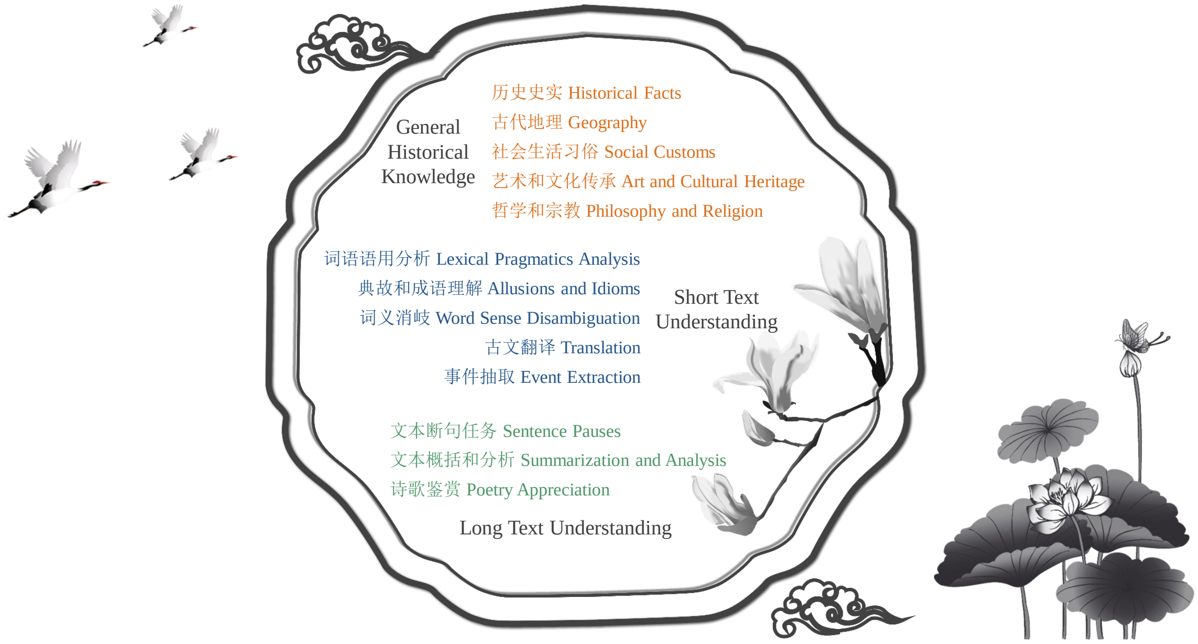

General Historical Knowledge
历史史实 Historical Facts
古代地理 Geography
社会生活习俗 Social Customs
艺术和文化传承 Art and Cultural Heritage
哲学和宗教 Philosophy and Religion
词语语用分析 Lexical Pragmatics Analysis
典故和成语理解 Allusions and Idioms
词义消岐 Word Sense Disambiguation
古文翻译 Translation
事件抽取 Event Extraction
Short Text Understanding
文本断句任务 Sentence Pauses
文本概括和分析 Summarization and Analysis
诗歌鉴赏 Poetry Appreciation
Long Text Understanding

Figure 1: Overview of AC-EVAL.

proficiency in ancient Chinese language understanding and historical knowledge. AC-EVAL comprises 3,245 multiple-choice questions, spanning three distinct dimensions and thirteen subjects, covering historical periods from the Pre-Qin to the Qing dynasty. These tasks, which progressively increase in difficulty, are categorized into general historical knowledge, short text understanding, and long text understanding. The general historical knowledge tasks address a diverse range of contents, including but not limited to, ancient historical facts, geography, social customs, art, religion and philosophy. Short text understanding covers lexical semantics and pragmatics, allusions and idioms, sentence translations, and event extraction. Long text understanding tasks focus on long text pauses, classical prose summarization and analysis, and the appreciation of themes, emotions and styles in poetry.

In our evaluation of LLMs on the AC-EVAL benchmark across answer-only (AO) and chain-of-thought (CoT) settings in zero- and few-shot scenarios, only ERNIE-Bot 4.0 and GLM-4 with accuracies over 70%. Results reveal significant improvement potential, especially in long text comprehension. Our analysis shows that Chinese LLMs outperform English ones in ancient Chinese. This distinction underscores the unique challenge that ancient Chinese as a low-resource area for models like GPT-4, despite their commendable performance on other Chinese benchmarks. Moreover, the broad range of knowledge required in our tasks reveals that LLMs encounter difficulties in grasping underlying rules, affecting few-shot learning outcomes. Interestingly, zero-shot CoT shows an advantage in larger models, underscoring the value of reasoning steps for complex tasks. Through the AC-EVAL benchmark, our goal is to provide a multidimensional evaluation tool, highlighting potential improvement areas to advance the development of LLMs in the understanding and education of ancient Chinese.

2 Related Work

2.1 Chinese benchmarks for LLMs

In the evolving landscape of NLP, the development of benchmarks to evaluate LLMs in comprehending Chinese has been a focal point of recent research (Chang et al., 2023). Benchmarks such as MMCU (Zeng, 2023), C-Eval (Huang et al., 2023), and CMMLU (Li et al., 2024) derived primarily from official examination questions, spanning various disciplines including STEM, humanities, social sciences, and professional qualification tests for fields like law and medicine. These benchmarks aimed to comprehensively assess the breadth of domains relevant to the Chinese language, primarily utilizing multiple-choice questions as their core components. Among these, CMMLU includes a small portion of ancient Chinese categories, approximately 0.7K, which cover topics like word-level semantics, historical facts, and modern Chinese literature. However, it lacks a broad temporal scope and content diversity, such as examinations of ancient geography, art and cultural heritage, and religion. Additionally, it does not focus on long texts, such as ancient Chinese reading comprehension. AGIEval (Zhong et al., 2023) expanded upon these by incorporating fill-in-the-blank questions alongside multiple-choice. CG-Eval (Zeng et al., 2023) and CLEVA (Li et al., 2023), on the other hand, took a more holistic approach to measure models' generative abilities, including tasks such as noun explanation, short answer questions, and computational problems. SuperCLUE (Xu et al., 2023) evaluated models across three dimensions: foundational abilities, professional knowledge, and Chinese language characteristics by leveraging actual user queries and ratings, along with a mix of open- and closed-ended questions. Lastly, OpenCompass (Contributors, 2023) integrates over 100 public datasets into a unified leaderboard framework, standardizing the assessment of LLMs.

Despite the extensive range of current benchmarks, there is a significant gap in their coverage of the ancient Chinese language, literature, and history. Considering the depth and breadth of Chinese millennia-long history, which includes evolving social customs, religious beliefs, geographical boundaries, and linguistic changes, it is evident that a more comprehensive benchmark is necessary.

2.2 Ancient Chinese benchmarks

Ancient Chinese, a fundamental component of the Chinese linguistic heritage, encapsulates millennia of historical narratives and cultural wisdom. A multitude of traditional and diverse datasets has been proposed to evaluate the ancient Chinese language understanding capabilities with various specific tasks (Pan et al., 2022; Wang and Ren, 2022; Liu et al., 2022; Tang and Su, 2022).

For instance, analyzing the sentiments and themes in poetry (as seen in FSPC (Shao et al., 2021) and TCCP (Liu et al., 2020), to the intricate task of translating between classical and modern Chinese, (illustrated by the Classical-Modern corpus[3] and the Erya dataset (Guo et al., 2023)). Furthermore, named entity recognition and relationship extraction tasks, with datasets like C-CLUE (Ji et al., 2021) and GuNER 2023[4], provide a foundation for in-depth linguistic analysis within ancient texts. GuwenEE[5], an event extraction dataset, is annotated and constructed from the "Twenty-Four Histories," a collection of Chinese official historical literature. The word sense disambiguation dataset for ancient Chinese, introduced by Shu et al. (2021), encompasses texts from multiple dynasties. Additionally, the EvaHan series from 2022 to 2024 introduces a spectrum of tasks including sentence segmentation, POS tagging, and machine translation. Comprehensive benchmarks like CCLUE and WYWEB (Zhou et al., 2023) integrate a variety of language understanding tasks, ranging from text classification to poetry analysis and machine reading comprehension, offering a holistic evaluation of models' linguistic proficiency.

However, despite the breadth of these benchmarks, there remains a discernible gap in the assessment of models' grasp of the historical knowledge hidden within ancient texts. The varied formats of datasets, designed for specific tasks, hinder uniform LLM evaluation, highlighting the urgent need for an integrated benchmark to thoroughly assess LLMs' understanding of ancient Chinese literature and history knowledge.

3 AC-EVAL Overview

3.1 Design Principles

The motivation behind constructing AC-EVAL is to comprehensively assess LLMs' understanding and reasoning capabilities regarding the shifts in societal customs, culture, and language throughout millennia of history. It adheres to four foundational principles to ensure a holistic evaluation framework:

Temporal Coverage: It spans from the pre-Qin period to the Qing dynasty, offering a broad historical scope that covers thousands of years of evolution. For the pre-Qin period, our dataset ranges from China's primitive society era (e.g., the Three Sovereigns and Five Emperors) to the slave dynasty period, including the Xia, Shang, and Zhou dynasties, roughly covering the period from 5000 BC to 221 BC. The Qing dynasty lasted from AD 1616 to 1912.

Task Difficulty Diversity: The benchmark ranges from basic fragmented historical knowledge to complex tasks requiring the understanding of ancient Chinese texts of various lengths, providing a graded evaluation of model capabilities. Our task

---

[3] https://github.com/NiuTrans/Classical-Modern

[4] https://guner2023.pkudh.org

[5] https://github.com/Lyn4ever29/GuwenEE

Category	Difficulty	# Subjects	# Questions	Average Length
General Historical Knowledge	Easy	5	1014	62.78
Short Text Understanding	Normal	5	1215	214.19
Long Text Understanding	Hard	3	1016	536.95

Table 1: Statistics of AC-EVAL. The average length is measured in characters.

difficulty classification is based on the actual characteristics of the tasks, expert advice, and results from preliminary small-scale tests. For example, tasks on historical knowledge mainly test memory and knowledge reserves, hence defined as easier tasks; short text comprehension tasks involve understanding vocabulary and sentences in ancient texts, considered to have medium difficulty; long text comprehension tasks require deeper context analysis and reasoning, therefore classified as more difficult tasks.

Content Diversity: It encompasses a broad spectrum of knowledge areas including historical facts, geography, religion, philosophy, social customs, architecture, music, and handicrafts, along with tasks in ancient language understanding such as semantic and syntactic analysis.

Data Quality: While ensuring the authority of the data, we also take specific measures to mitigate data contamination, as detailed in section 3.2.

Our benchmark is organized into 3 major categories and 13 subjects, encompassing general historical knowledge as well as both short- and long-text comprehension of ancient Chinese. In alignment with the methodology proposed by Huang et al. (2023), we adopt a uniform question format, presenting each question with four answer options. Each subject within the benchmark contains an average of over 200 questions, of which five with explanations are designated for development sets. The statistical summary of AC-EVAL is depicted in Table 1, and a more detailed statistical breakdown is available in Appendix A.

3.2 Data Collection

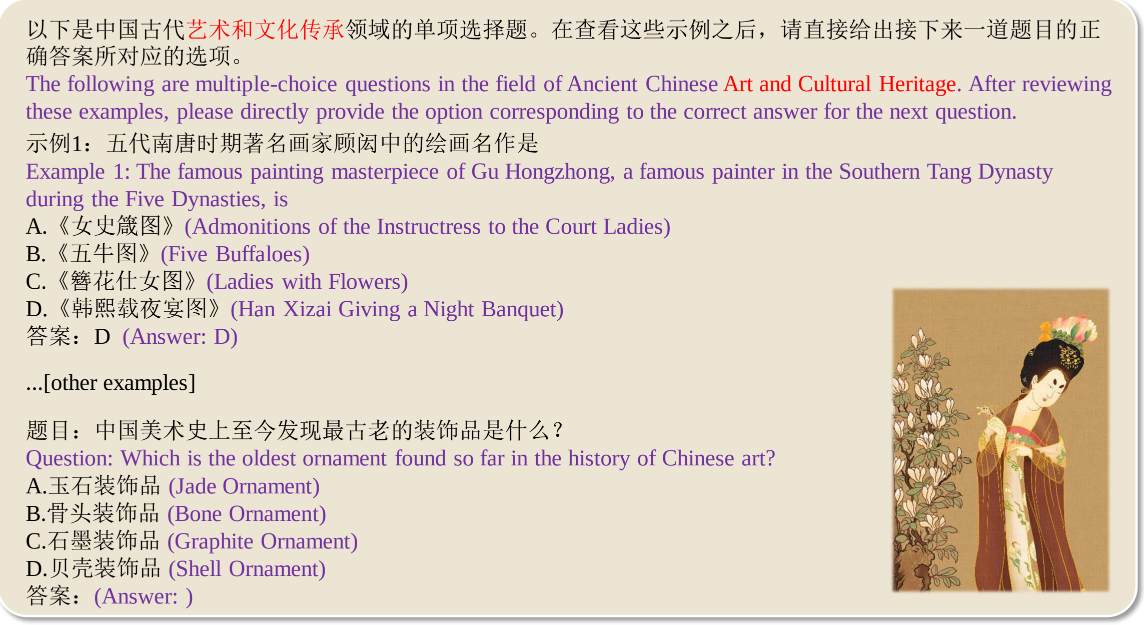
Subject Selection: Our benchmark encompasses general historical knowledge and ancient Chinese text comprehension. For the former, we have identified five subcategories, namely: Historical Facts, Geography, Social Customs, Art and Cultural Heritage, Philosophy and Religion. For the latter, we distinguish between short texts, which include tasks such as Lexical Pragmatics Analysis, Allusions and Idioms, Word Sense Disambiguation, Translation, and Event Extraction, and long texts, which cover Sentence Pauses, Summarization and Analysis, and Poetry Appreciation, as illustrated in Figure 1.

Data Source: The dataset is derived from four main sources: (1) the Complete Library in Four Branches (Siku Quanshu), offering a comprehensive collection of ancient Chinese texts; (2) specialized books on ancient Chinese social customs, architectural history, music history, and geography; (3) official or mock examinations; and (4) existing non-multiple-choice datasets on ancient Chinese, such as GuwenEE. Further details on the data sources are provided in Appendix A.

Data processing: Initially, we recruited undergraduate students and linguistics experts as annotators to manually gather and compile preliminary questions and answers from these sources. The data then underwent a three-fold modification and review process: **(1) Ethical Considerations:** We categorized our data source into reference materials (Sources 1, 2, and 4) and examinations (Source 3). The reference materials were manually adapted to create new questions and answers. Meanwhile, all materials are cited appropriately in Appendix A. The examination data, available freely online, were also included. **(2) Data Contamination:** We aimed to strike a balance between maintaining the authority of the data sources and minimizing data contamination. With the awareness that official examinations might be inadvertently captured and utilized in training LLMs, these were adapted by experts to retain the examinations' core focus while altering the content to some extent. **(3) Coverage and Accuracy:** We adhered strictly to our design principles to ensure the dataset's diversity and accuracy. 5% random sample of the data underwent a quality check, with any found inaccuracies necessitating rework until achieving 100% accuracy. For a more detailed expert evaluation process, please refer to Appendix B.

3.3 Evaluation

Accuracy is the primary metric for our evaluation. The ground-truth labels of the development set are



以下是中国古代艺术和文化传承领域的单项选择题。在查看这些示例之后，请直接给出接下来一道题目的正确答案所对应的选项。
The following are multiple-choice questions in the field of Ancient Chinese Art and Cultural Heritage. After reviewing these examples, please directly provide the option corresponding to the correct answer for the next question.
示例1：五代南唐时期著名画家顾闳中的绘画名作是
Example 1: The famous painting masterpiece of Gu Hongzhong, a famous painter in the Southern Tang Dynasty during the Five Dynasties, is
A.《女史箴图》(Admonitions of the Instructress to the Court Ladies)
B.《五牛图》(Five Buffaloes)
C.《簪花仕女图》(Ladies with Flowers)
D.《韩熙载夜宴图》(Han Xizai Giving a Night Banquet)
答案：D (Answer: D)

...[other examples]

题目：中国美术史上至今发现最古老的装饰品是什么？
Question: Which is the oldest ornament found so far in the history of Chinese art?
A.玉石装饰品 (Jade Ornament)
B.骨头装饰品 (Bone Ornament)
C.石墨装饰品 (Graphite Ornament)
D.贝壳装饰品 (Shell Ornament)
答案：(Answer:)

Figure 2: Illustrative few-shot AO prompts from AC-EVAL with corresponding English translations for better readability.

public, while the labels of the test set remain confidential to avoid their unintended inclusion in the pre-training corpora. For a more detailed evaluation process, please refer to our GitHub link.

4 Experiment

4.1 Setup

For evaluation of the AC-EVAL benchmark, we assess LLMs in both zero-shot and few-shot settings, with the few-shot samples drawn from the development set. To extract the answer choices from the models' responses, We employ regular expressions followed by manual verification to ensure successful retrieval in nearly all cases.

We report the results for both answer-only (AO) and chain-of-thought (CoT) (Wei et al., 2022b; Dong et al., 2022; Zhang et al., 2022) settings in zero- and few-shot scenarios. For zero-shot AO setting, we craft prompts in the format: "以下是中国古代[主题]领域的单项选择题，请直接给出正确答案对应的选项。(The following is a multiple-choice question in the field of Ancient Chinese [subject]. Please directly provide the option corresponding to the correct answer.)" For the few-shot AO setting, an example of it prompt is displayed in Figure 2. The logic behind our selection of few-shot examples is similar to the principles of data construction, requiring broad coverage across various dynasties and diverse content topics. For instance, in the case of the five examples related to arts and cultural heritage, we provide content related to calligraphy, architecture, painting, sculpture, etc., from different historical periods. For the CoT settings, their prompts are shown in Appendix C.

Generally, few-shot defaults to five-shot. It is noteworthy that for both the five-shot and five-shot-CoT settings, input lengths sometimes surpass the maximum token limit of the models. To accommodate this, we dynamically adjust the number of samples to ensure they fit within the models' context window constraints.

4.2 Models

In our evaluation, we select 17 top-performing LLMs that demonstrate proficiency in Chinese language comprehension. These models represent a variety of organizations and encompass a range of parameter sizes. For commercial models, we evaluate via API calls, including (1) GPT-4 and GPT-3.5 Turbo (Achiam et al., 2023), (2) ERNIE-Bot 4.0 and ERNIE-Bot[6], (3) GLM-4 and GLM-3-Turbo (Zeng et al., 2022), (4) Qwen-max (Bai et al., 2023). For models with open-sourced parameters, we evaluate (1) LLaMA2-70B (Touvron et al., 2023) (2) Qwen-7B/14B/72B-Chat (Bai et al., 2023), (3) Yi-6B/34B-Chat[7], (4) Baichuan2-

---

[6] https://cloud.baidu.com/
[7] https://huggingface.co/01-ai

Model	General Historical Knowledge	Short Text Understanding	Long Text Understanding	Average
GPT-4	66.11	55.11	47.38	56.20
GPT-3.5 Turbo	53.50	43.72	36.94	44.72
ERNIE-Bot 4.0	**77.54**	**68.11**	66.42	**70.69**
ERNIE-Bot	68.81	57.80	51.47	59.36
GLM-4	76.63	66.66	**67.70**	70.33
GLM-3-Turbo	75.21	60.52	59.77	65.17
Qwen-max	73.77	64.88	63.84	67.50
LLaMA2-70B	33.55	36.29	30.72	33.54
Qwen-72B-Chat	71.25	61.48	59.80	64.18
Yi-34B-Chat	72.66	61.33	58.36	64.12
Qwen-14B-Chat	69.51	56.53	57.38	61.14
Baichuan2-13B-Chat	65.57	49.24	35.40	50.07
Qwen-7B-Chat	62.74	48.76	44.97	52.16
Baichuan2-7B-Chat	64.38	46.77	40.33	50.49
Yi-6B-Chat	66.70	47.79	39.49	51.33
ChatGLM3-6B	58.04	43.01	39.73	46.93
Xunzi-Qwen-Chat	60.20	44.31	30.87	45.13

Table 2: Zero-shot AO average accuracy of all models. We report average accuracy over subjects within each category. “Average” = average over all categories. Models are ranked by model size.

7B/13B-Chat (Baichuan, 2023), (5) ChatGLM3-6B[8], and (6) Xunzi-Qwen-Chat[9]—an LLM that has been continually pre-trained on ancient Chinese corpora based on the Qwen-7B-Chat architecture. A detailed description of the evaluated models, including their architectural details, pretraining corpora, and versions, is available in Appendix D. We conduct timely evaluations to capture the latest performance levels of these models[10].

5 Results

In this section, we explore the comparative performance of various models under four distinct settings: zero-shot AO as discussed in Section 5.1, few-shot AO in Section 5.2, zero- and few-shot CoT in Section 5.3.

5.1 Zero-shot AO

Given that zero-shot scenarios are among the most common use cases, understanding model performance in this context is crucial. Therefore, we first report the average accuracy in the zero-shot AO setting in Table 2, while detailed accuracy breakdowns by subject are provided in Appendix E. Our comparison analysis focuses on two critical dimensions: model parameter size and task category.

Comparison by model. For large models: ERNIE-Bot-4.0 and GLM-4 stand out as top-performing models in ancient Chinese, with accuracies of 70.69% and 70.33%, respectively, followed by Qwen-max at 67.50%. Despite primarily being trained on modern Chinese corpora, these LLMs show strong generalization abilities to ancient Chinese. For models primarily trained on English corpora, GPT-4 and GPT-3.5 significantly outperform LLaMA-70B. Considering our benchmark is entirely in Chinese, this suggests GPT models’ superior generalization capabilities over LLaMA2-70B in handling extensive Chinese content. Interestingly, GPT series models perform worse than Chinese LLMs, diverging from conclusions drawn from previous benchmarks in the Chinese domain where GPT often ranked first (Li et al., 2024; Huang et al., 2023; Xu et al., 2023). This indicates that ancient Chinese acts as a low-resource language for English LLMs, highlighting the significant linguistic differences between ancient and modern Chinese. This observation also underscores the importance of our benchmark from another perspective.

For small models: The Yi-34B-Chat showcases remarkable parameter efficiency and performs com-

---

[8] https://github.com/THUDM/ChatGLM3

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

[10] All models were evaluated during 5-10 February 2024.

Model	General Historical Knowledge	Short Text Understanding	Long Text Understanding	Average
GPT-4	65.91 (-0.20)	58.07 (+2.96)	48.36 (+0.98)	57.45 (+1.25)
GPT-3.5 Turbo	53.99 (+0.49)	43.21 (-0.51)	36.40 (-0.54)	44.54 (-0.18)
ERNIE-Bot 4.0	**75.69** (-1.85)	**69.59** (+1.48)	66.12 (-0.30)	**70.47** (-0.22)
ERNIE-Bot	68.81 (+0.00)	57.62 (-0.18)	50.36 (-1.11)	58.93 (-0.43)
GLM-4	74.89 (-1.74)	65.48 (-1.18)	**69.07** (+1.37)	69.81 (-0.52)
GLM-3-Turbo	72.99 (-2.22)	59.48 (-1.04)	59.66 (-0.11)	64.04 (-1.13)
Qwen-max	75.29 (+1.52)	65.48 (+0.60)	66.99 (+3.15)	69.25 (+1.75)
Qwen-72B-Chat	71.67 (+0.42)	61.30 (-0.18)	57.07 (-2.73)	63.35 (-0.83)
Yi-34B-Chat	66.62 (-6.04)	52.57 (-8.76)	41.90 (-16.46)	53.70 (-10.42)
Qwen-14B-Chat	70.60 (+1.09)	53.73 (-2.80)	45.91 (-11.47)	56.75 (-4.39)
Baichuan2-13B-Chat	63.75 (-1.82)	45.86 (-3.38)	32.74 (-2.66)	47.45 (-2.62)
Qwen-7B-Chat	61.42 (-1.32)	45.98 (-2.78)	30.78 (-14.19)	46.06 (-6.10)
Baichuan2-7B-Chat	63.37 (-1.01)	45.91 (-0.86)	39.94 (-0.39)	49.74 (-0.75)
Yi-6B-Chat	55.76 (-10.94)	35.97 (-11.82)	28.48 (-11.01)	40.07 (-11.26)
ChatGLM3-6B	55.74 (-2.30)	42.92 (-0.09)	38.45 (-1.28)	45.71 (-1.22)
Xunzi-Qwen-Chat	51.30 (-8.90)	41.25 (-3.06)	29.84 (-1.03)	40.80 (-4.33)

Table 3: Few-shot AO average accuracy of all models. We report average accuracy over subjects within each category. "Average" = average over all categories. The values in parentheses show the relative change compared to the zero-shot AO scenario.

parably to larger models like Qwen-72B-Chat. This efficiency can be attributed to their extensive training on large-scale Chinese corpora and architectural optimizations. Qwen-14B even surpasses the GPT series and ERNIE-bot, presenting high cost-effectiveness in ancient Chinese comprehension as a relatively smaller open-source model. Qwen-7B-Chat achieves the best performance among models with less than 10B parameters. Baichuan2-13B-Chat, compared to Baichuan2-7B-Chat, does not show performance improvement despite increased parameters, possibly due to a reduction in focus on ancient Chinese content in its training corpus. Xunzi-Qwen-Chat, despite being fine-tuned on ancient Chinese texts, shows a decline in performance compared to Qwen-7B-Chat. This highlights the trade-off between specialized knowledge and general applicability.

Comparison by Task Category. (1) General Historical Knowledge: Most models score highest on this category of tasks, likely because these tasks focus on the retrieval and understanding of factual information without necessitating deep textual analysis and reasoning. (2) Short Text Understanding: Compared to long text comprehension, models generally score higher on short text understanding, though still lower than on general historical knowledge tasks. This may be because short text understanding still requires models to capture subtle semantic differences and contextual relationships, albeit with relatively lower complexity. (3) Long Text Comprehension: All models generally score lower on long text comprehension than on other tasks, indicating it as a challenging task that requires advanced understanding, reasoning, and synthesis capabilities.

5.2 Few-shot AO

Table 3 presents the results for the few-shot AO setting, alongside a comparison with the zero-shot AO setting.

For large models, only GPT-4 and Qwen have a 1-2% improvement in this setting, while others slightly declined, diverging from previous Chinese benchmarks where few-shot usually excelled over zero-shot (Huang et al., 2023). We attribute this discrepancy primarily to the task specificity. Previous benchmarks often encompassed a broader range of subjects and task categories, including scientific, technological, and coding tasks. In such tasks, few-shot learning effectively aids models in capturing underlying patterns, thereby enhancing adaptability and generalization. However, our tasks focus on a broad spectrum of fragmented knowledge and require a deep understanding of ancient Chinese, including its cultural, historical backgrounds, and linguistic structures, leading to a unique challenge where few-shot learning might

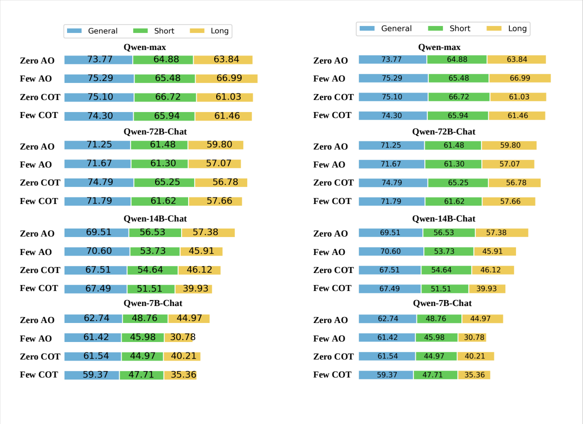


Figure 3: Average accuracy for each category in various settings with different model sizes. Here, we have omitted the category names and shot types for brevity.

not provide the same level of benefit. Instead, the specificity and complexity of ancient Chinese tasks can result in the introduction of noise or irrelevant information through few-shot examples, potentially hindering model performance rather than enhancing it.

For small models, the decline in performance is more pronounced. The Yi-series models, although performing comparably to large models in the zero-shot AO scenario, show the most significant drop in the few-shot setting, with a performance decrease of 10%. This aligns with conclusions from some previous studies (Li et al., 2024), suggesting that for smaller models, few-shot learning may introduce too much irrelevant content, potentially leading to information interference. In such instances, models might struggle to extract useful knowledge or patterns from a few samples, as the additional information introduced may not be entirely relevant to the task, thereby diluting the model's focus.

From this analysis, we conclude that despite the task specificity causing few-shot learning to sometimes act as interference, large models possess stronger language understanding capabilities and higher stability in processing distracting information, even achieving improvements on some tasks. However, smaller models struggle due to insufficient parameters to effectively encode and utilize contextual information.

5.3 Chain-of-thought

As shown in Figure 3, we conduct a series of experiments to explore the impact of CoT on LLMs of varying parameter sizes. Our experimental setup includes two scenarios: zero-shot CoT and few-shot CoT (prompts are detailed in Appendix C). To facilitate a nuanced comparison, we select the Qwen series, encompassing four different sizes: 7B, 14B, 72B, and the non-public Qwen-max. Given that few-shot performance does not surpass zero-shot on AC-EVAL, our analysis primarily contrasts the two CoT formats against the zero-shot AO scenario.

Zero-shot CoT vs. Zero-shot AO: In zero-shot CoT, prompts are adjusted to encourage stepwise analysis. This method particularly benefits large models like Qwen-max and Qwen-72B-Chat in historical knowledge and short text understanding tasks but shows a decrease in long text comprehension. We attribute this to the increased reasoning steps needed for long text understanding in large models, where any small errors can accumulate, negatively impacting the final answer's accuracy.

As model size decreases, a downward trend in performance is evident across all tasks in the zero-shot CoT setting. This decline is likely due to the CoT method's demand for models to understand the question, generate intermediate reasoning steps, and ultimately formulate an answer. This process, more complex than direct answer generation, requires robust semantic understanding and logical reasoning capabilities. With reduced model parameters, the capability of models to perform these functions weakens, leading to diminished performance.

Few-shot CoT vs. Zero-shot AO/CoT: Few-shot CoT underperforms in comparison to zero-shot settings in both AO and CoT across all parameter sizes. This aligns with our above observation that few-shot learning generally offers less benefit in our benchmark, which demands a broad understanding of fragmented knowledge and deep comprehension of ancient Chinese, including its cultural, historical backgrounds, and linguistic structures. The unique challenges posed by these requirements suggest that even structured CoT, when combined with few-shot examples, may be perceived as informational noise, thereby impeding the model's ultimate reasoning capability.

Through this analysis, it is evident that while CoT reasoning can enhance model performance in certain contexts, the effectiveness of this approach is contingent upon the model's capacity for complex information processing and logical deduction. The decline in performance with reduced model size and the limited impact of few-shot learning highlight the intricate balance required between model abilities, task specificity, and the introduced format of CoT.

6 Conclusion

We introduce AC-EVAL, a benchmark designed to evaluate LLMs' proficiency in ancient Chinese, addressing a gap by covering historical knowledge and language understanding extensively. Our experiments reveal significant improvement areas for existing LLMs. We identify critical factors influencing LLM performance and suggest practical directions for enhancing these models. AC-EVAL aims to advance LLM application in ancient Chinese education, offering a valuable tool for assessing and developing Chinese LLMs.

7 Limitations

While our study introduces the AC-EVAL benchmark as a robust tool for evaluating LLMs in the domain of ancient Chinese, it is imperative to acknowledge several limitations that accompany our research:

Absence of Human Baseline: The lack of a human comparative standard impedes the evaluation of LLMs' depth of understanding, cultural acuity, and contextual sensitivity relative to the insights provided by scholars specializing in ancient Chinese literature. Consequently, while the AC-EVAL benchmark may offer quantitative evaluations of LLM proficiency, it might not capture the qualitative dimensions of linguistic and cultural comprehension that are crucial in the analysis of ancient Chinese texts.

Focus on Multiple-Choice Questions: The current iteration of AC-EVAL primarily utilizes a multiple-choice format to assess LLMs. This approach, while effective in certain assessments, does not measure the generative capabilities of LLMs. For example, poetry generation (Chen et al., 2019; Zhipeng et al., 2019). As a result, our benchmark may not fully capture the models' ability to produce coherent and contextually relevant responses in an open-ended format.

In light of these limitations, future work will aim to incorporate human evaluation and expand the benchmark to include open-ended and generative tasks, thereby enhancing the comprehensive assessment of models' capabilities.

Acknowledgments

This work is supported by the National Natural Science Foundations of China under Grant (61972047, 62372060).

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*.

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*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4925–4931.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Geyang Guo, Jiarong Yang, Fengyuan Lu, Jiaxin Qin, Tianyi Tang, and Wayne Xin Zhao. 2023. Towards effective ancient chinese translation: Dataset, model, and evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 416–427. Springer.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Zijing Ji, Yuxin Shen, Yining Sun, Tian Yu, and Xin Wang. 2021. C-clue: a benchmark of classical chinese based on a crowdsourcing system for knowledge graph construction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 295–301. Springer.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11260–11285.

Yanyang Li, Jianqiao Zhao, Duo Zheng, Zi-Yuan Hu, Zhi Chen, Xiaohui Su, Yongfeng Huang, Shijia Huang, Dahua Lin, Michael R. Lyu, and Liwei Wang. 2023. Cleva: Chinese language models evaluation platform. *arXiv preprint arXiv:2308.04813*.

Yuqing Li, Yuxin Zhang, Bin Wu, Ji-Rong Wen, Ruihua Song, and Ting Bai. 2022. A multi-modal knowledge graph for classical Chinese poetry. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2318–2326.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Maofu Liu, Junyi Xiang, Xu Xia, and Huijun Hu. 2022. Contrastive learning between classical and modern chinese for classical chinese machine reading comprehension. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–22.

Yutong Liu, Bin Wu, and Ting Bai. 2020. The construction and analysis of classical chinese poetry knowledge graph (in Chinese). *Journal of Computer Research and Development*, 57(6):1252.

Xiaomeng Pan, Hongfei Wang, Teruaki Oka, and Mamoru Komachi. 2022. Zuo zhuan Ancient Chinese dataset for word sense disambiguation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 129–135.

Yizhan Shao, Tong Shao, Minghao Wang, Peng Wang, and Jie Gao. 2021. A sentiment and style controllable approach for chinese poetry generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 4784–4788.

Lei Shu, Yiluan Guo, Huiping Wang, Xuetao Zhang, and Renfen Hu. 2021. The construction and application of Ancient Chinese corpus with word sense annotation. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 549–563.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Xuemei Tang and Qi Su. 2022. That slepen al the nyght with open ye! cross-era sequence segmentation with switch-memory. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7830–7840.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Pengyu Wang and Zhichen Ren. 2022. The uncertainty-based retrieval framework for Ancient Chinese CWS and POS. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168.

Ping Wang, Shitou Zhang, Zuchao Li, and Jingrui Hou. 2023. Enhancing Ancient Chinese understanding with derived noisy syntax trees. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 83–92.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*.

Hui Zeng, Jingyuan Xue, Meng Hao, Chen Sun, Bin Ning, and Na Zhang. 2023. Evaluating the generation capabilities of large chinese language models. *arXiv preprint arXiv:2308.04823*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

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*.

Guo Zhipeng, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative Chinese classical poetry generation system. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 25–30.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP evaluation benchmark for classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Details of AC-EVAL

Table 5 provides a comprehensive overview of the AC-EVAL, detailing the data sources and the specific concepts addressed within each subject. Table 6 offers insights into the quantitative aspects of the dataset, including the number of questions per subject and their average length (accounting for both the questions and explanations, measured in characters). Furthermore, Table 4 shows the distribution of choices across the multiple-choice questions.

Option	C-EVAL	AC-EVAL
A	22.9%	26.2%
B	26.0%	26.5%
C	26.4%	23.6%
D	24.7%	23.7%

Table 4: Distribution of Answers

B Construction process of AC-EVAL

In the process of developing a high-quality dataset for ancient Chinese natural language understanding tasks, a systematic approach is adopted that encompasses several key steps. These steps ensure the rigorous collection, annotation, and evaluation of data, ultimately leading to a robust and reliable dataset. The following outlines the major phases of this process:

Task Collection Experts systematically collect common tasks related to ancient Chinese, creating a clear list of tasks that are categorized into knowledge-based and language understanding categories to better meet research needs.

Purpose and Principle For different subjects, experts establish detailed purposes and principles for data collection and annotation, specifying requirements for time coverage, content diversity, and annotation consistency to ensure high-quality and representative data.

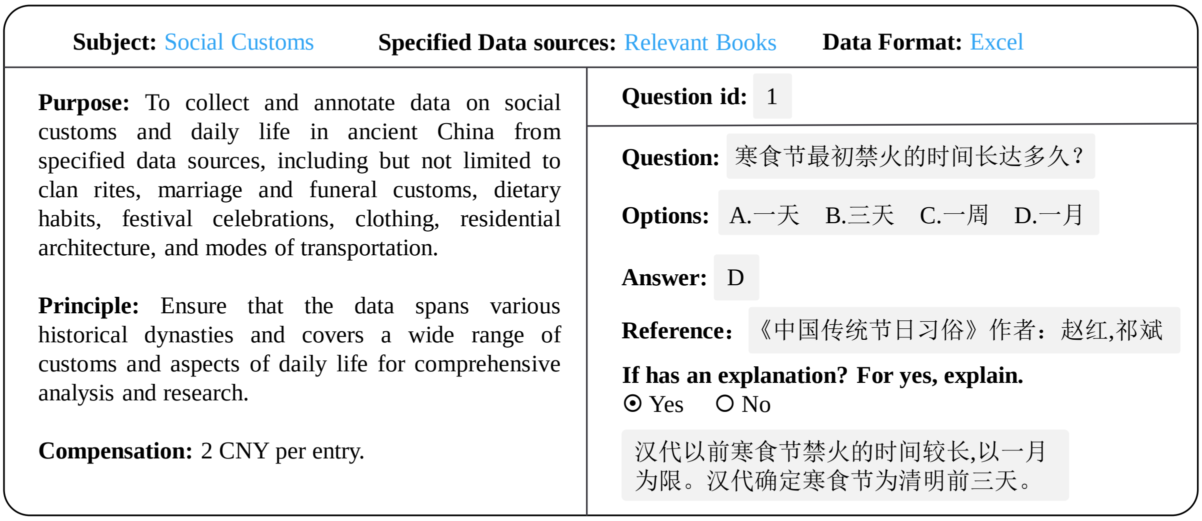
Data Collection and Annotation Data is collected and annotated manually for each task category. Figure 4 details the specific requirements for data annotation. All our annotators are Chinese undergraduate students and experts in Chinese linguistics, and they are compensated at a rate that meets market standards. Each data entry includes a question, four answer options, and the corresponding correct answer, striving to cover a wide range of scenarios and ensuring data comprehensiveness and accuracy.

Quality Check Experts conduct random sampling of the annotated data, typically reviewing 5% of the samples for verification. If errors are found, feedback is provided to the relevant standard developers for a comprehensive data review until a 100% pass rate is achieved, ensuring data quality.

Task Difficulty Division After ensuring data verification, small sample testing is performed to evaluate performance in a large model. Experts classify questions into easy, normal, and hard categories based on multiple dimensions, including question difficulty, model output content, and scoring results, facilitating subsequent analysis and application.

C Prompts for Evaluation

Figures 5 and 6 display the chain-of-thought evaluation prompts used in the zero-shot and few-shot settings, respectively.



Subject: Social Customs **Specified Data sources:** Relevant Books **Data Format:** Excel

Purpose: To collect and annotate data on social customs and daily life in ancient China from specified data sources, including but not limited to clan rites, marriage and funeral customs, dietary habits, festival celebrations, clothing, residential architecture, and modes of transportation.

Principle: Ensure that the data spans various historical dynasties and covers a wide range of customs and aspects of daily life for comprehensive analysis and research.

Compensation: 2 CNY per entry.

Question id: 1

Question: 寒食节最初禁火的时间长达多久？

Options: A.一天 B.三天 C.一周 D.一月

Answer: D

Reference: 《中国传统节日习俗》作者：赵红,祁斌

If has an explanation? For yes, explain.
● Yes ○ No

汉代以前寒食节禁火的时间较长,以一月为限。汉代确定寒食节为清明前三天。

Figure 4: Illustration of the Annotation Process: An Example of Social Customs Data Annotation

D Details of the LLMs being evaluated

We provide a detailed description of the Large Language Models (LLMs) that were rigorously evaluated during the period of 5-10 February 2024, ensuring the assessment of the latest model versions prior to submission.

GPT (Achiam et al., 2023) series models, developed by OpenAI, designed to be more aligned with human-like interaction, exhibiting helpful, safe, and truthful behavior as enhanced by Reinforcement Learning from Human Feedback (RLHF). GPT-4, with its ability to process images, PDFs, and other file types, underwent a comprehensive post-training alignment process. We evaluate the versions of gpt-3.5-turbo-0125 and gpt-4-0125-preview.

ERNIE-Bot is an industrial-grade, knowledge-enhanced LLM developed by Baidu. The 4.0 version represents a significant upgrade in understanding, generation, logic, and memory capabilities over its predecessors, supporting extensive input and output lengths (5K input + 2K output). Our evaluation included both ERNIE-Bot and ERNIE-Bot 4.0.

GLM (Zeng et al., 2022) series, developed by Zhipu AI and Tsinghua University, are bidirectional dense models excelling in bilingual language processing. ChatGLM, a derivative of GLM, targets Chinese QA and dialogue tasks with enhanced fine-tuning and feedback. We evaluate ChatGLM3-6B and the commercial GLM-3-Turbo and GLM-4.

Qwen (Bai et al., 2023), developed by Alibaba, is trained on a vast corpus including 3 trillion tokens of texts and codes. The chat variants of Qwen have been refined through RLHF to better align with human preferences. We conduct a comprehensive evaluation of the Qwen series, covering multiple versions with varying parameter sizes, including Qwen-7B/14B/72B-Chat and Qwen-max.

LLaMA2 (Touvron et al., 2023), developed and open-sourced by Meta AI, excels in encoding, inference, and knowledge application. It incorporates several enhancements over the vanilla Transformer architecture employed by preceding LLMs, optimizing for greater training efficiency. In our experiment, we evaluate the performance of the LLaMA2-70B version.

Yi series models by 01.AI are open-source bilingual models trained from scratch on a 3T multilingual corpus, featuring an extended context window of up to 200K tokens. We utilize the Yi-6B-Chat and Yi-34B-Chat versions, which support up to 32K tokens for context in inferences.

Baichuan2 (Baichuan, 2023) is developed by Baichuan Intelligence Inc., trained on a 2.6 trillion token high-quality corpus and supporting multiple languages including Chinese, English and others. The versions evaluated are Baichuan2-7B-Chat and Baichuan2-13B-Chat.

Xunzi is a model collaboratively released by Nanjing Agricultural University and the Zhonghua Book Company. It is fine-tuned on ancient Chinese corpora such as the Siku Quanshu, based on foundations from Qwen, Baichuan, and GLM. We evaluate the Xunzi-Qwen-Chat, a model trained from Qwen-7B-Chat.

E Breakdown of Model Performance

Table 7 provides a detailed accuracy breakdown by subject for four representative models under AO settings in both zero- and few-shot scenarios, respectively. Comprehensive results for all models are made available on GitHub.

F Error Analysis

We conduct an in-depth error analysis across different tasks using specific examples. Under the zero-shot setting, we compare four models from the Qwen series: Qwen-7B-Chat, Qwen-14B-Chat, Qwen-72B-Chat, and Qwen-max. Our objective is to evaluate their capabilities and identify areas for improvement.

Historical Knowledge LLMs often exhibit excellent performance due to their extensive parameters and rich knowledge bases. As shown in Table 8, we focused on tasks related to ancient geography. The 7B and 14B models provided incorrect answers, while the 72B model answered correctly but lacked adequate explanations. Qwen-Max offered explanations but included some erroneous information. This indicates that both large and small models have deficiencies in knowledge explanation for ancient Chinese, especially in historical contexts.

Short Text Understanding Using the lexical pragmatics analysis task as an example (see Table 9), the smaller models (7B and 14B) made errors in explaining and classifying special usages of words in ancient Chinese texts. In contrast, the larger models (72B and Qwen-Max) were able to grasp and explain these special usages more accurately. This suggests that larger model parameters may contribute to better understanding and analysis of texts with complex semantics and structures. However, errors still exist in larger models. For instance, in option C provided by Qwen-Max, the word "至" is identified as a verb meaning "to arrive," but it should refer to "those who arrived." Only 72B correctly recognized this subtle grammatical nuance.

Long Text Understanding We examined the poetry appreciation task (see Table 10) to assess how models handle long texts with complex literary and historical backgrounds. Smaller models struggled, often misinterpreting symbols and metaphors within the texts. Larger models demonstrated a better ability to comprehend deeper meanings and cultural implications but still require improvements in precision and consistency. Notably, larger models like Qwen-Max performed better in handling complex texts, largely due to their enhanced ability to understand literary and historical contexts. This capability allows these models not only to recognize the direct meaning of the text but also to capture deeper symbolism and cultural significance.

Overall, LLMs still have significant shortcomings in understanding ancient Chinese language. One major issue is the hallucination of historical knowledge and the mishandling of deep literary and historical content, which affects their reliability in educational applications. Integrating more suitable Retrieval-Augmented Generation (RAG) methods presents a promising avenue for innovation. By enhancing models with accurate and context-specific external knowledge, we can improve their ability to handle complex historical and literary tasks, thereby increasing their trustworthiness in educational settings. Furthermore, the prompts currently used are generally generic, lacking specificity for different tasks. Designing task-specific and guiding prompts for various applications, and training models with these tailored prompts, can enable them to learn how to process ancient Chinese texts more effectively.

Additionally, given the high computational costs associated with large-parameter models, exploring model distillation techniques to transfer the capabilities of large models to smaller ones is a feasible research direction. This approach can reduce dependence on large-scale computational resources while maintaining or even enhancing performance levels. Training a smaller model (with 7B or 14B parameters) to support comprehensive knowledge understanding is a practical and valuable goal for future research. This not only makes the technology more accessible but also allows for broader applications in resource-constrained environments, ultimately advancing the field of natural language understanding in ancient Chinese.

Subject	**Data Source**	**Concepts**
Historical Facts	Official history exams	Historical facts, covering political, economic, and military developments across different periods.
Geography	10% from mock exams and 90% from ancient place names knowledge database	Administrative divisions, historical boundaries,changes in place names over time.
Social Customs	Relevant Books , e.g., Customs of the Qing Dynasty.	Changes in clothing, food, housing, transportation, traditional festivals, weddings and funerals, family etiquette, public customs, business practices, entertainment customs over time.
Art and Cultural Heritage	Mock exams and relevant books, e.g., History of Chinese Art, Architecture and Music History	Changes in calligraphy, painting, architecture, craftsmanship and music over different periods.
Philosophy and Religion	Mock exams and relevant books e.g., History of Chinese philosophy.	Changes in the content of Taoism, Confucianism, Buddhism, etc., and their rise and decline over time.
Lexical Pragmatic Analysis	Compiled by linguistics experts	Flexible usage of parts of speech and figures of speech.
Allusions and Idioms	Official and mock exam questions	Allusions and idioms and the cultural meaning behind them.
Word Sense Disambiguation	Word Sense Disambiguation Dataset (Shu et al., 2021)	Explanation of word meaning in a given text.
Translation	Classical-Modern Chinese translation dataset[11]	Overall understanding of the semantics and syntax of sentences.
Event Extraction	GuwenEE[12]	Identifying basic facts and information in short texts, such as time, location, characters, event types, etc.
Sentence Pauses	Siku Quanshu	Make pauses in reading unpunctuated ancient writings.
Summarization and Analysis	Official and mock exams	Overall understanding, analysis, and reasoning for ancient Chinese texts
Poetry Appreciation	Official and mock exams	Analysis of imagery, style, sentiment in classical Chinese poetry

Table 5: Data Sources and Concepts for All Subjects.

Subject	Test		Dev		
	# Questions	Len. of Q	# Questions	Len. of Q	Len. of E
Historical Facts	199	157.1	5	138.0	200.2
Geography	197	33.8	5	32.8	33.6
Social Customs	202	48.5	5	48.6	65.0
Art and Cultural Heritage	195	35.8	5	32.4	56.8
Philosophy and Religion	196	39.2	5	48.0	77.4
Lexical Pragmatic Analysis	198	62.5	5	69.6	75.4
Allusions and Idioms	206	191.2	5	79.6	132.4
Word Sense Disambiguation	402	176.6	5	163.2	91.4
Translation	199	409.1	5	315.0	79.4
Event Extraction	185	238.8	5	150.4	109.0
Sentence Pauses	202	390.2	5	404.2	294.2
Summarization and Analysis	598	880.5	5	856.0	341.4
Poetry Appreciation	201	339.4	5	371.8	109.0

Table 6: Quantitative Statistics for All Subjects.

以下是中国古代艺术和文化传承领域的单项选择题，请逐步分析并给出正确答案对应的选项。
The following is a multiple-choice question in the field of Ancient Chinese Art and Cultural Heritage. Please analyze step by step and provide the option corresponding to the correct answer.

题目：中国美术史上至今发现最古老的装饰品是什么？
Question: Which is the oldest ornament found so far in the history of Chinese art?
A.玉石装饰品 (Jade Ornament)
B.骨头装饰品 (Bone Ornament)
C.石墨装饰品 (Graphite Ornament)
D.贝壳装饰品 (Shell Ornament)
答案：(Answer:)

Figure 5: Illustrative zero-shot CoT prompts from AC-EVAL with corresponding English translations for better readability.

以下是中国古代艺术和文化传承领域的单项选择题。在查看这些示例之后，请逐步分析接下来一道题目并给出正确答案所对应的选项。
The following are multiple-choice questions in the field of Ancient Chinese Art and Cultural Heritage. After reviewing these examples, please analyze the next question step by step and provide the option corresponding to the correct answer.

示例1：五代南唐时期著名画家顾闳中的绘画名作是
Example 1: The famous painting masterpiece of Gu Hongzhong, a famous painter in the Southern Tang Dynasty during the Five Dynasties, is
A.《女史箴图》(Admonitions of the Instructress to the Court Ladies)
B.《五牛图》(Five Buffaloes)
C.《簪花仕女图》(Ladies with Flowers)
D.《韩熙载夜宴图》(Han Xizai Giving a Night Banquet)
答案：
让我们逐步分析。顾闳中的绘画名作是《韩熙载夜宴图》。《五牛图》是韩滉的作品，《簪花仕女图》是周昉的作品，《女史箴图》是顾恺之的作品。
所以答案是D。
Answer:
Let‘s analyze step by step. The famous painting by Gu Hongzhong is 'Han Xizai Giving a Night Banquet.' 'Five Buffaloes' is a work by Han Huang, 'Ladies with Flowers' is by Zhou Fang, and 'Admonitions of the Instructress to the Court Ladies' is by Gu Kaizhi.
Therefore, the answer is D.

...[other examples]

题目：中国美术史上至今发现最古老的装饰品是什么？
Question: Which is the oldest ornament found so far in the history of Chinese art?
A.玉石装饰品 (Jade Ornament)
B.骨头装饰品 (Bone Ornament)
C.石墨装饰品 (Graphite Ornament)
D.贝壳装饰品 (Shell Ornament)
答案：(Answer:)

Figure 6: Illustrative few-shot CoT prompts from AC-EVAL with corresponding English translations for better readability.

Subject	ERNIE-Bot 4.0	GLM-4	Yi-34B-Chat	Qwen-7B-Chat
Historical Facts	78.39/76.76	78.89/84.42	75.88/58.29	61.81/61.31
Geography	78.17/74.11	75.13/75.13	71.57/75.63	66.50/62.44
Social Customs	79.21/78.22	77.23/75.74	76.73/70.79	70.79/71.29
Art and Cultural Heritage	74.87/72.82	76.92/75.90	70.26/68.21	59.49/61.03
Philosophy and Religion	77.04/76.53	75.00/63.27	68.88/60.20	55.10/51.02
Lexical Pragmatics Analysis	78.22/83.17	75.25/79.70	77.72/57.43	44.55/27.23
Allusions and Idioms	66.67/70.20	58.59/59.60	48.48/44.95	33.84/32.83
Word Sense Disambiguation	67.96/70.39	69.90/68.93	70.39/64.08	59.71/52.43
Translation	56.22/56.22	59.45/52.49	49.50/35.82	39.80/36.57
Event Extraction	71.86/74.37	71.86/72.86	65.33/47.74	47.74/43.22
Sentence Pauses	56.86/52.01	59.20/59.36	42.64/32.44	37.63/31.77
Summarization and Analysis	64.18/63.18	68.66/68.16	54.73/35.82	52.74/33.33
Poetry Appreciation	77.84/76.76	73.51/73.51	72.97/70.27	62.70/64.86

Table 7: Accuracy per subject in the answer-only setting: a comparison of zero-shot (left) and few-shot (right) performance.

Example 1	古地名"澶州"这一名字，在历史上首次出现于哪个朝代? The ancient place name "Chanzhou" first appeared in history during which dynasty? A. 南宋 (Southern Song) B. 北宋 (Northern Song) C. 唐 (Tang) D. 东汉 (Eastern Han) 答案：C (Answer: C)
Explanation	隋开皇十六年（公元596年），置澶渊县。唐初，为避唐高祖李渊讳改名澶水县。唐武德四年（621）置澶州。 (In the 16th year of the Kaihuang era of the Sui Dynasty (AD 596), Chanyuan County was established. In early Tang, to avoid the personal name taboo of Emperor Gaozu of Tang, Li Yuan, it was renamed Chanshui County. In the 4th year of the Wude era of the Tang Dynasty (AD 621), Chanzhou Prefecture was established.)
Qwen-7B-Chat	...澶州这个名字首次出现在北宋时期。... (...The name "Chanzhou" first appeared during the Northern Song period...)
Qwen-14B-Chat	...北宋时期（960年-1127年），曾设置过澶州，这是它首次出现在历史上的记载。... (...During the Northern Song period (960–1127 AD), Chanzhou was established, marking its first appearance in historical records...)
Qwen-72B-Chat	根据历史记载，"澶州"这个名字最早出现在唐朝时期。... (According to historical records, the name "Chanzhou" first appeared during the Tang Dynasty...)
Owen-max	...根据历史记载，澶州在唐初为澶水县，唐玄宗开元二十二年（734年）设置澶州，因此"澶州"这一地名首次出现于唐朝。... (...According to historical records, Chanzhou was initially Chanshui County in early Tang. In the 22nd year of the Kaiyuan era under Emperor Xuanzong of Tang (AD 734), Chanzhou Prefecture was established. Therefore, the place name "Chanzhou" first appeared during the Tang Dynasty...)

Table 8: Error analysis for general historical knowledge.

Example 2	下列各组句子中[]内的词用法不同于其他的一项是： Which of the words in brackets has a different usage compared to the others? A. [歌]窈窕之章 ([Singing] articles related to the bright moon.) B. 假舟楫者，非能[水]也 (Those who use a boat cannot [swim].) C. 则其[至]又加少矣 (Therefore, its [arrival] is even less.) D. 唐浮图慧褒始[舍]于其址 (The Tang Dynasty's stupa Huibao began to [settle] at its site.) 答案：C (Answer: C)
Explanation	A.名词作动词，唱；B.名词作动词，游水；C.动词作名词，到达的人；D.名词作动词，筑舍定居。 (A. Noun used as a verb, meaning "to sing"; B. Noun used as a verb, meaning "to swim"; C. Verb used as a noun, meaning "the one who arrives"; D. Noun used as a verb, meaning "to build a shelter and settle down".)
Qwen-7B-Chat	...我们逐一分析每个选项。A. "歌"在此处是一个动词，指唱歌或请求歌唱，与其他三个选项不同。B."水"在这里也是动词，表示游泳，与其他选项相同。C. "至"在这里也是动词，表示到达，与其他选项相同。D."舍"在这里也是动词，表示筑舍定居，与其他选项相同。综上所述，答案为A。 (Let's analyze each option one by one. A. "歌" here is a verb, meaning "to sing," which differs from the other three options. B. "水" here is also a verb, meaning "to swim," similar to the others. C. "至" here is also a verb, meaning "to arrive," similar to the others. D. "舍" here is also a verb, meaning "to build a shelter and settle down," similar to the others. In summary, the answer is A.)
Qwen-14B-Chat	...[水]和[舍]在古文中都为名词作动词的用法，表示"游泳"和"建造"的意思。[至]在古文中则是形容词作名词的用法，表示"到达的人数"。而[歌]则是在这里作为动词，表示歌唱的意思，与其他三个词不同。综上所述，本题答案选A... (In classical Chinese, [水] and [舍] are both nouns used as verbs, meaning "to swim" and "to build," respectively. In classical Chinese, [至] is an adjective used as a noun, meaning "the number of arrivals." However, [歌] is used as a verb here, meaning "to sing," which is different from the other three words. In summary, the answer to this question is A...)
Qwen-72B-Chat	...A项："歌"在此处是动词，意为"歌唱"；B项："水"在此处也是动词，意为"游泳"；C项："至"在此处是名词，意为"到达的人"；D项："舍"在此处是动词，意为"筑舍定居"。综上所述，本题答案选C... (A: "歌" here is a verb, meaning "to sing"; B: "水" here is also a verb, meaning "to swim"; C: "至" here is a noun, meaning "the one who arrives"; D: "舍" here is a verb, meaning "to build a shelter and settle down." In summary, the answer to this question is C...)
Owen-max	A选项中，"歌"是名词活用为动词，意为"吟诵"；B选项中，"水"是名词活用为动词，意为"游泳"；C选项中，"至"是动词，意为"到"；D选项中，"舍"是名词活用为动词，意为"筑舍居住"。通过分析可以发现，C选项中的"至"没有发生词类活用，用法不同于其他项。因此，正确答案是C。 (In option A, "歌" is a noun used as a verb, meaning "to recite;" In option B, "水" is a noun used as a verb, meaning "to swim;" In option C, "至" is a verb, meaning "to arrive;" (In option D, "舍" is a noun used as a verb, meaning "to build a shelter and live in." Through analysis, it can be found that "至" in option C has not undergone a change in word class, making its usage different from the others. Therefore, the correct answer is C.)

Table 9: Error analysis for short text understanding.

Example 3	...多景楼中昼掩扉，画梁不敢住乌衣。 禅房花木兵烧杀，佛寺干戈僧怕归。 山雨欲来淮树立，潮风初起海云飞。 酒尊未尽登舟急，更过金焦看落晖。...下列对这首诗的理解和赏析，不正确的一项是（ ） (...Even in Duojing Tower, I have to close the doors and windows in broad daylight. Even in the richly decorated Wuyi Alley, I dare not stay overnight. The once tranquil meditation rooms and temples, now after the ravages of war, are still filled with warfare and withered flowers and trees; even the monks living there are afraid to return. The mountain rain is about to come, the trees on the banks of the Huai River stand tall; the tidal wind has just risen, clouds fly over the vast water surface. Before finishing the wine in my cup, I hurriedly boarded the boat, once again facing the boundless sunset alone at Jinshan and Jiaoshan... Among the following understandings and appreciations of this poem, which one is incorrect?) A. 白日闭门，诗人所见多景楼已非昔时繁华。战乱驱使人群四散，首句中通过""画梁""、""乌衣""的意象运用了借代手法以展现这一景象。(Closing doors in broad daylight shows that Duojing Tower seen by the poet is no longer as prosperous as before. War drives people to scatter; in the first sentence, the imagery of "painted beams" and "Wuyi" uses metonymy to present this scene.) B.元军侵袭南方，肆虐破坏，即使是佛家清修的寺庙禅林也未能避免其害，"僧怕归"这一描绘直接而具体地展现了该情境。(The Yuan army invaded the south, wreaking havoc; even the pure temples and meditation forests of Buddhism could not avoid harm. The depiction of "monks afraid to return" directly and concretely presents this situation.) C.自然界中的""山雨欲来""与""潮风初起""预示着社会动荡不安的局势，诗人刚从惊恐中稍定，又因战事突发，匆匆登上船只逃离。(The natural phenomena of "mountain rain about to come" and "tidal wind just rising" foreshadow the turbulent social situation. The poet, just slightly calmed from fear, hurriedly boarded the boat to flee due to the sudden outbreak of war.) D. 面对国土丧失、战火弥漫，诗人心中满溢悲愤与无奈。回望西沉落日余晖，那既是皇朝衰败的形象象征，亦是诗人内心深切悲痛情感的折射镜像。 (Facing the loss of the country and pervasive war, the poet's heart is filled with grief and helplessness. Looking back at the setting sun in the west, it is both a symbolic representation of the dynasty's decline and a reflection of the poet's deep sorrowful emotions.) 答案：B (Answer: B)
Explanation	B选项理解错误，"僧怕归"是通过僧人的心理状态侧面反映出战乱的场景，而非直接描述。 (Option B is incorrectly understood; "monks afraid to return" indirectly reflects the scene of war through the psychological state of the monks, rather than direct description.)
Qwen-7B-Chat	...选项C则错误。"山雨欲来淮树立，潮风初起海云飞"这两句描绘了自然界即将发生的变化，预示着将要到来的战事。并没有描述诗人刚刚从惊恐中恢复过来。...综上所述，本题答案选C. (...Option C is incorrect. The lines "mountain rain about to come, trees stand by the Huai River; tidal wind just rising, clouds fly over the sea" depict upcoming changes in nature, foreshadowing the impending war. There is no description of the poet just recovering from fear. ...In summary, the answer to this question is C.)
Qwen-14B-Chat	...选项A中，"白日闭门"确实表现出战乱导致人们生活的不稳定和恐惧，但"画梁"和"乌衣"并非借代手法。"画梁"指的是装饰华丽的屋梁，"乌衣"则是指黑色的衣服，这两个词在这里都是具体的实物描写...因此，选项A是错误的。综上所述，本题答案选A... (...In Option A, "closing doors in broad daylight" indeed shows the instability and fear in people's lives caused by war, but "painted beams" and "Wuyi" are not using metonymy. "Painted beams" refers to gorgeously decorated beams, "Wuyi" refers to black clothes; these two terms here are concrete descriptions... Therefore, Option A is incorrect. In summary, the answer is A...)
Qwen-72B-Chat	...B项，根据"兵烧杀"、"僧怕归"可以推断出，"僧怕归"并不是直接描写，而是通过对僧人的描绘间接展现出战乱的情况；C项，"山雨欲来"与"潮风初起"是自然界的现象，而非社会动荡不安的局势；...综上所述，本题答案选B。(...For Option B, based on "soldiers burning and killing" and "monks afraid to return," we can infer that "monks afraid to return" is not a direct description but indirectly reflects the war situation through the depiction of monks; For Option C, "mountain rain about to come" and "tidal wind just rising" are natural phenomena, not the turbulent social situation; ...In summary, the answer is B.)
Owen-max	...B选项分析："僧怕归"一词是通过僧人的心理状态侧面反映出战乱对佛寺的破坏和影响，而非直接具体描绘情境。故B选项错误。...综上所述，本题答案选B。(...Analysis of Option B: The phrase "monks afraid to return" reflects the destruction and impact of war on Buddhist temples through the psychological state of monks, rather than directly and concretely depicting the situation. Therefore, Option B is incorrect. ...In summary, the answer is B.)

Table 10: Error analysis for long text understanding.