

Can LLMs Act as Historians? Evaluating Historical Research Capabilities of LLMs via the Chinese Imperial Examination

Lirong Gao^{1,3*}, Zeqing Wang^{2,3*}, Yuyan Cai^{2,3*}, Jiayi Deng^{3*},
Yanmei Gu³, Yiming Zhang^{1†}, Jia Zhou^{2†}, Yanfei Zhang^{2†}, Junbo Zhao^{1,3†}

¹State Key Laboratory of Blockchain and Data Security, Zhejiang University

²Zhejiang University ³Ant Group

{gaolirong, 12343006, 22443022, yimingz, 0012802, yanfei.zhang, j.zhao}@zju.edu.cn,
{dengjiayi.djy, guyanmei.gym}@antgroup.com

Abstract

While Large Language Models (LLMs) have increasingly assisted in historical tasks such as text processing, their capacity for professional-level historical reasoning remains underexplored. Existing benchmarks primarily assess basic knowledge breadth or lexical understanding, failing to capture the higher-order skills—such as evidentiary reasoning—that are central to historical research. To fill this gap, we introduce **ProHist-Bench**, a novel benchmark anchored in the *Chinese Imperial Examination (Keju)* system—a comprehensive microcosm of East Asian political, social, and intellectual history spanning over 1,300 years. Developed through deep interdisciplinary collaboration, ProHist-Bench features 400 challenging, expert-curated questions across eight dynasties, accompanied by 10,891 fine-grained evaluation rubrics. Through a rigorous evaluation of 18 LLMs, we reveal a significant proficiency gap: even state-of-the-art LLMs struggle with complex historical research questions. We hope ProHist-Bench will facilitate the development of domain-specific reasoning LLMs, advance computational historical research, and further uncover the untapped potential of LLMs. We release ProHist-Bench at <https://github.com/inclusionAI/ABench/tree/main/ProHist-Bench>.

1 Introduction

Recent breakthroughs in LLMs (OpenAI et al., 2024; Yang et al., 2025) profoundly impact historical research by automating the processing of vast digitized archival collections (Chen et al., 2025) and facilitating narrative generation (Ye et al., 2025). Such paradigm shift motivates the development of historical benchmarks for LLMs. Existing evaluations cover various dimensions, including general historical knowledge (Hauser et al., 2024),

*Equal contribution

†Corresponding authors

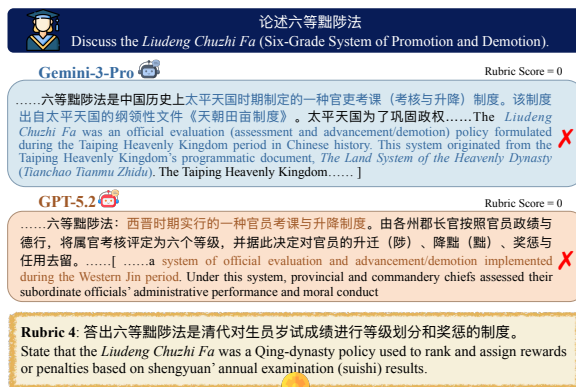


Figure 1: An illustration of LLMs' factual hallucinations in historical research tasks. Severe hallucinations persist in specialized historical contexts, even among advanced LLMs.

ancient languages (Zhou et al., 2023; Piryani et al., 2024), and multi-modal historical materials (Shi et al., 2023; Ghaboura et al., 2025; Zhang et al., 2025b), providing the evaluation of LLMs' comprehension of general history and their ability to process complex historical materials.

However, a significant gap remains between basic historical knowledge and professional historical research. As shown in Figure 1, LLMs still suffer from hallucination problems and fail to resolve conflicting historical records, which are also demonstrated in recent studies (Zhou et al., 2025; Ghaboura et al., 2025). Rather than merely retrieving facts, professional historical research hinges on complex skills such as socio-political contextualization, evidentiary reasoning and interpretive synthesis (Weinstein, 2005; Chen, 2014). Yet current benchmarks largely overlook the evaluation of such higher-order skills, leaving LLM capabilities on genuinely professional historical tasks both underexplored and easy to overestimate.

To bridge this gap, we propose **ProHist-Bench**, the first **Professional History Benchmark** framework. Through deep interdisciplinary collaboration

between AI researchers and professional historians, ProHist-Bench is anchored in the *Chinese Imperial Examination* (*Keju*, 科举), a civil service selection system that operated for over 1,300 years and profoundly influenced the East Asian civilizations (Weinstein, 2005; Chang, 1942). Centered on *Keju*, historians curated 400 expert-level, challenging questions and, crucially, handcrafted fine-grained scoring rubrics tailored to each question, amounting to 10,891 rubrics in total. These criteria systematically evaluate 9 core historical research capabilities (e.g., evidentiary reasoning, temporal reframing), ensuring comprehensive and rigorous professional assessment. Our systematic evaluation and analysis of 18 advanced LLMs reveals significant deficiencies in professional historical research capabilities, identifying specific weaknesses in handling complex, conflicting, and context-dependent historical tasks.

In summary, our contributions are as follows:

- We construct 400 challenging history questions written by historians, and for the first time use the *Keju* as a distinctive lens to cover roughly 1,300 years of historical evolution in ancient China, forming a multi-level benchmark that spans from basic factual understanding to historical research tasks.
- We further propose the first history evaluation framework for LLMs: historians manually created 10,891 scoring rubrics covering nine categories of historical research abilities, and we also designed a complete rubric-based evaluation pipeline to enable systematic and reproducible assessment of model performance.
- We conducted a systematic evaluation of 18 advanced LLMs. Our results reveal significant deficiencies in current models' ability to perform professional historical research.

2 Related Work

2.1 MLLM Evaluation for History

In the multimodal domain, researchers have established benchmarks evaluating cross-cultural understanding (Liu et al., 2021; Chiu et al., 2025; Vayani et al., 2024), Chinese visual culture (Zhang et al., 2025a; Zhou et al., 2025), and historical material processing (Shi et al., 2023; Chen et al., 2025; Ghaboura et al., 2025; Liu et al., 2025; Zhang et al., 2025b). However, these benchmarks primarily focus on basic perceptual tasks such as visual recognition and OCR, largely neglecting the deep

reasoning and argumentation capabilities essential for professional historical research.

2.2 LLM Evaluation for History

A burgeoning body of evaluation research has explored the application of LLMs within historical research. Existing work covers global history knowledge (**HiST-LLM** (Hauser et al., 2024), **HiBenchLLM** (Chartier et al., 2025)), as well as historical document understanding (**M5HisDoc** (Shi et al., 2023), **ChroniclingAmericaQA** (Pirayani et al., 2024), **AC-EVAL** (Wei et al., 2024), and **C³Bench** (Cao et al., 2024b)). For **general Chinese capabilities**, benchmarks such as **C-Eval** (Huang et al., 2023), **CMMLU** (Li et al., 2023), and **Chinese SimpleQA** (He et al., 2024) provide comprehensive assessments of fundamental knowledge and QA across diverse disciplines. In the specialized field of **Classical Chinese Studies (CCS)**, **ACLUE** (Zhang and Li, 2023) and **C-CLUE** (Ji et al., 2021) target tasks ranging from ancient text comprehension to Named Entity Recognition, while **WYWEB** (Zhou et al., 2023) and **WenMind** (Cao et al., 2024a) expand the scope to include sequence labeling, machine translation, and holistic alignment with human intuition. While existing benchmarks effectively assess general knowledge and linguistic proficiency, they fail to capture specialized capabilities such as low-resource context adaptation and deep historical reasoning.

Compared with existing work, the uniqueness and advantages of ProHist-Bench are twofold. First, ProHist-Bench is the first benchmark focused on in-depth research into the history of *Keju*, moving beyond basic historical QA toward specialized historical research. Second, we establish a more rigorous and fine-grained evaluation framework for assessing research-oriented abilities, including fact organization, historical comparison, and evidentiary reasoning. This framework also provides a valuable reference for future benchmark construction in other historical domains.

3 Dataset Construction

3.1 Design Principles

Historical research presents distinctive reasoning challenges that extend beyond basic knowledge comprehension to encompass accurate terminology interpretation, evidence-based deduction, and the construction of coherent narratives grounded in historical context. ProHist-Bench aims to bridge

Task ID	Task Name	Task Description
T1	Term Interpretation	Definition: Assesses the model’s ability to understand and explain historical terms. Example: Explain the term <i>Gongshi</i> (graduate of the <i>Keju</i> examination).
T2	Fact QA	Definition: Assesses the model’s ability to organize and present historical facts. Example: Briefly explain how the Qing dynasty’s entry inspection system prevented cheating in the <i>Keju</i> examination.
T3	Historical Reasoning	Definition: Assesses the model’s ability to make comparisons, integrate viewpoints, and reason with evidence based on historical facts. Example: Discuss the <i>Liudeng Chuzhi Fa</i> (Six-Grade System of Promotion and Demotion).
T4	<i>Celun</i> Generation	Definition: Assesses the model’s cross-contextual reasoning and decision-making abilities within a specific historical context. Example: Assume you are a Qing dynasty candidate in the <i>Keju</i> examination in the 46th year of Qianlong (1749). Please write an essay in the Eight-Legged Essay (baguwen) format, following the specific rules for writing. If any characters need to be avoided due to taboos, use pinyin. Limit your response to 700 words. Provide only the essay and do not include any other content. The topic is: 孟子曰：待文王而后兴者，凡民也 (Mencius said: Those who wait for Wenwang to rise are all the people.).

Table 1: Task definitions in ProHist-Bench. The benchmark is organized into four task types: Term Interpretation (T1), Fact QA (T2), Historical Reasoning (T3), and *Celun* Generation (T4).

the gap between comprehensive LLM evaluation and the intrinsic challenges of historical research. Aligned with the characteristics of historical scholarship, we adhere to the following principles in constructing ProHist-Bench.

Historical Representativeness. We emphasize that historians are typically not generalists, but rather experts in a particular historical field, focusing on in-depth research of specific historical periods or events. The *Keju* examinations, as an important mechanism for talent selection in ancient China, spans over 1,300 years of history, and its study represents an extremely important direction for contemporary Chinese historical research. Our dataset contain a substantial amount of historical knowledge about *Keju*, and only by mastering this knowledge can historical researchers conduct in-depth research on *Keju* and thereby provide new insights. Although *Keju* research differs from other historical research fields in terms of research content, their research methodologies often share similarities. Therefore, examining LLMs’ capabilities in fact organization, evidence reasoning, and other aspects through this topic can authentically reflect the performance level of LLMs when conducting in-depth historical research.

Historical Authority. Historical authority is not an inherent or immutable property of any single narrative, but rather a defensible interpretation grounded in available evidence and established

scholarly norms. Accordingly, drawing on a substantial body of extant historical materials and historiographical scholarship—such as *Qingdai Zhujuan Jicheng* (清代朱卷集成) and the latest academic findings—we develop a rigorously grounded and well-justified evaluation framework to assess the capabilities of LLMs.

Multi-dimensional Capability Assessment. A well-designed benchmark should capture the diverse capabilities required for historical research. To this end, ProHist-Bench is constructed along four dimensions. First, in terms of **task diversity**, as shown in Table 1, ProHist-Bench draws inspiration from historical examinations and covers four task types: Term Interpretation (T1), Fact QA (T2), Historical Reasoning (T3), and *Celun* (Policy Essay) Generation (T4). Second, in terms of **difficulty diversity**, all questions are annotated as either *General* or *Hard*, depending on whether they primarily require basic factual understanding or more advanced abilities such as multi-step reasoning, multi-source integration, and interpretive judgment. This binary scheme reduces subjective ambiguity and improves annotation consistency. Third, in terms of **content diversity**, ProHist-Bench spans a broad range of historical domains, including politics, economics, social life, transportation, and intellectual-cultural history. Fourth, in terms of **temporal coverage**, it covers more than 1,300 years across nine dynasties, enabling evaluation across varied historical periods.

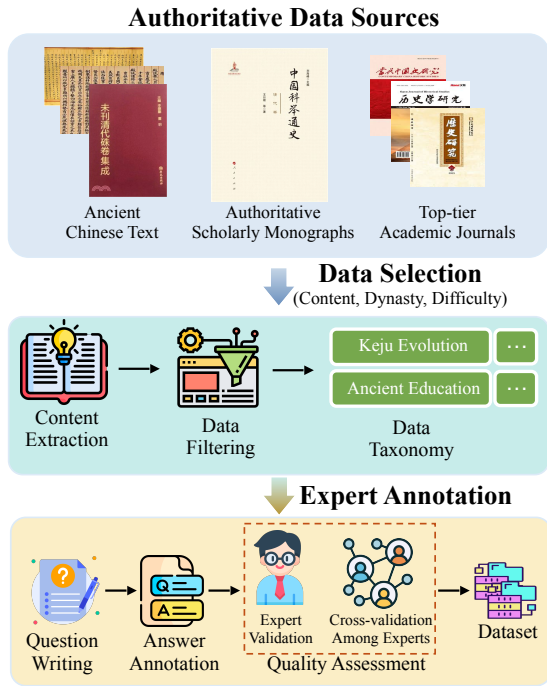


Figure 2: Overview of the dataset construction pipeline.

3.2 Authoritative Data Sources

ProHist-Bench is constructed based on 125 core references spanning three categories (see Appendix H for the complete bibliography): **(1) Ancient Chinese Text.** Qing dynasty official documents, local gazetteers, and examination archives, including *Qingdai Zhujuan Jicheng* (清代朱卷集成), and *Qingdai Keju Kaoshi Shulu* (清代科举考试述录) by the last *Tanhua* (探花) (third-place scholar). **(2) Authoritative Scholarly Monographs.** Two Chinese national comprehensive histories, alongside seminal works by leading scholars including Ichisada Miyazaki, Benjamin Elman, and others, ensuring global scholarly perspectives. **(3) Top-tier Academic Journals.** Cutting-edge research papers from CSSCI-indexed journals, including *Historical Research*, *Modern Chinese History Studies*, and *Qing History Journal*, etc.

3.3 Expert Annotation

The questions and reference answers in this dataset were all manually constructed by historians. The dataset construction pipeline is shown in Figure 2.

Question Writing Based on thirteen preset dimensions and temporal distributions, we extracted related knowledge from historical materials and research papers, converting them into four question types: Term Interpretation, Fact QA, Historical Reasoning, and *Celun* Generation. We introduced

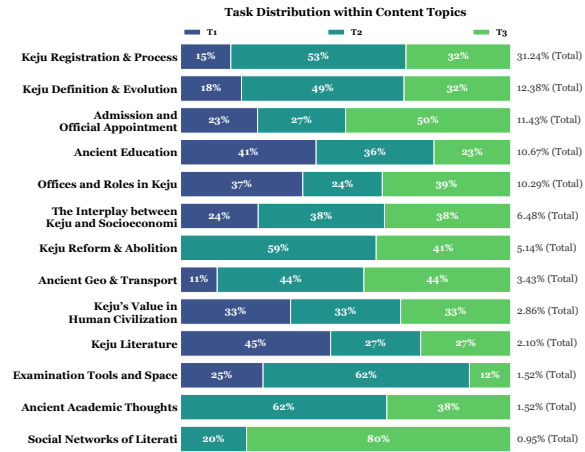


Figure 3: Distribution of task categories across various content topics. The chart displays the ratio of T1-T3 tasks for each topic.

LLMs to assist historians in refining the question phrasing: identifying ambiguities through multi-round Q&A testing, adjusting qualifiers and background prompts accordingly, and attempting different questioning styles to evaluate the models' diverse capabilities.

Answer Annotation We consulted a total of 125 references to compile the answers and provided detailed citations for each one. The answers underwent cross-verification via multi-source historical materials, adhering to mainstream consensus while listing various viewpoints for unsettled issues, emphasizing argumentative logic rather than trivial details. It is worth noting that, although we attempted to use LLMs to assist in answer annotation, we found that LLMs exhibit significant shortcomings in subject completion within ancient texts, terminology understanding, and the mechanisms of historical events. Therefore, ultimately, all reference answers were written sentence-by-sentence by historians to ensure historical accuracy.

Quality Assessment To ensure data quality, we adopt a four-stage quality-control pipeline. First, six historians independently wrote 100 questions each based on 125 literatures and annotated reference answers, rubrics, and metadata such as dynasty and topic. Second, each question was reviewed by at least one non-original historian, and any disagreements were adjudicated by a third senior historian, ensuring at least two-person review for every question. Third, two senior historians independently checked all questions, and any remaining disagreements were resolved by majority

ID	Capability Category	Count	Percentage
R1	Concept Definition	1,002	9.20%
R2	Fact Organization	6,452	59.24%
R3	Historical Comparison	173	1.59%
R4	Evidentiary Reasoning	732	6.72%
R5	Comprehensive Evaluation	315	2.89%
R6	Viewpoint Integration	97	0.89%
R7	Academic Expression	1,167	10.72%
R8	Classical Writing	91	0.84%
R9	Temporal Reframing	862	7.91%
Total	/	10,891	100%

Table 2: Statistics of specific rubric dimensions (R1–R9) in ProHist-Bench.

voting, providing full-coverage and high-standard quality control. Fourth, we randomly sampled 5% of the data for in-depth review by historians. If any error was found, a full recheck was triggered, followed by repeated sampling until the sampled batch reached 100% accuracy. In addition, we discarded questions that were too easy, overly subjective, or highly contested such that stable criteria could not be defined within the designated source boundary, thereby ensuring the executability and consistency of the evaluation standards.

3.4 Dataset Statistics

ProHist-Bench consists of 400 expert-curated historical research questions spanning four core tasks: Term Interpretation, Fact QA, Historical Reasoning, and *Celun* Generation. As shown in Figure 3, the dataset covers 13 topics, with Registration and Process accounting for the largest share (31.24%), followed by Definition and Evolution (12.38%) and Entry into Officialdom and Appointments (11.43%). In terms of task distribution, T2 constitutes the largest proportion (43%), followed by T3 (35%) and T1 (22%). More fine-grained statistics by task type, difficulty, and dynasty are provided in Appendix C.

4 Rubric Evaluation Framework

To systematically evaluate whether LLMs can function like historians, we designed a multi-dimensional rubric evaluation framework, decomposing LLM performance into 9 observable dimensions, defining clear levels and rules for each dimension, thereby enhancing scoring consistency and reproducibility.

4.1 Rubric Definition

A rubric is a set of criteria or scoring rules developed by historians to evaluate the quality of responses to historical questions. As shown in Table 2, we design nine rubric dimensions to capture different aspects of historical research ability. For T1–T3, the rubric covers **Concept Definition** (R1), **Fact Organization** (R2), **Historical Comparison** (R3), **Evidentiary Reasoning** (R4), **Comprehensive Evaluation** (R5), **Viewpoint Integration** (R6), and **Academic Expression** (R7). Together, these dimensions assess whether a model can accurately define historical concepts, organize and compare facts, reason with evidence, synthesize scholarly perspectives, and present arguments in a rigorous academic manner. For T4 (*Celun* Generation), we further introduce **Classical Writing** (R8) and **Temporal Reframing** (R9), which evaluate the model’s ability to generate historically grounded responses under dynasty-specific literary, institutional, and stylistic constraints. To preserve historical fidelity, the rubric also imposes explicit penalties for major violations of period conventions, such as departures from genre norms or the use of taboo expressions. Detailed definitions, scoring rules, and penalty criteria are provided in Appendix A.

Each dimension is weighted according to its importance for historical research and the level of capability it demands from the model. Basic dimensions, such as Concept Definition and Fact Organization, mainly assess knowledge recall and are therefore assigned lower weights. By contrast, more advanced dimensions, such as Evidentiary Reasoning and Temporal Reframing, require broader knowledge and more complex reasoning, and thus receive higher weights. The nine dimensions are applied selectively across tasks: T1–T3 emphasize R1–R7, while T4 focuses on the specialized abilities captured by R8 and R9.

Penalty Rubrics. In addition to the positive evaluation rubrics designed for the nine capabilities above, we further establish a set of universal penalty rubrics applicable to all tasks, aimed at capturing intolerable errors, hallucinations, and other problematic behaviors in LLM-based historical research. Detailed definitions are provided in Appendix B.

4.2 Rubric Construction

Constructing fine-grained and objective rubrics is challenging, but necessary for in-depth evaluation of LLMs in history. Existing benchmarks mainly rely on automated metrics such as Accuracy and BLEU, which capture surface-level textual similarity but fail to assess higher-order historical abilities, such as fact organization, comparative analysis, and evidence-based interpretation.

Rubric-based evaluation provides a more suitable alternative. While this approach has been validated in expert domains such as law (Shi et al., 2026) and healthcare (Arora et al., 2025), our work is, to the best of our knowledge, the first to apply it to history. All rubrics in our benchmark are question-specific and developed by historians. For T1–T3, we adopt an **Iterative Refinement** strategy, where historians draft initial criteria and iteratively revise them based on real model outputs to avoid overly coarse judgments. For T4, we follow a **Historical Authenticity** principle, grounding the rubrics in the *Qing Dynasty Imperial Examination Regulations* (清代科场条例) and relevant scholarship to ensure historical fidelity and temporal accuracy. To enhance objectivity, we invited multiple historians to translate historiographical requirements into quantifiable criteria and conducted repeated cross-validation, reducing individual bias and strengthening disciplinary rigor.

4.3 Rubric Statistics

As shown in Tables 2 and 3, ProHist-Bench comprises 10,891 fine-grained evaluation criteria, with an average of 27.23 criteria per question. T1–T3 tasks contain 9,938 criteria (91.25%), while T4 task include 953 criteria (8.75%). Among the nine capability categories, Fact Organization accounts for the largest proportion (59.24%), reflecting its fundamental role in historical research, followed by Academic Expression (10.72%) and Concept Definition (9.20%). The distribution demonstrates our rubric criteria emphasis on core historical competencies while maintaining comprehensive coverage of diverse historical research capabilities.

4.4 Evaluation Protocol

We adopt an LLM-as-a-Judge paradigm to automate the evaluation process. To ensure reliable evaluation, we conducted a preliminary study comparing six candidate LLMs against expert historian evaluations. We randomly sampled 50 instances

Task	Total # Criteria	Avg. # Criteria
T1	1,650	18.33
T2	3,963	26.42
T3	4,325	36.04
T4	953	23.83
Total	10,891	27.23

Table 3: Statistics of fine-grained evaluation criteria across different task categories.

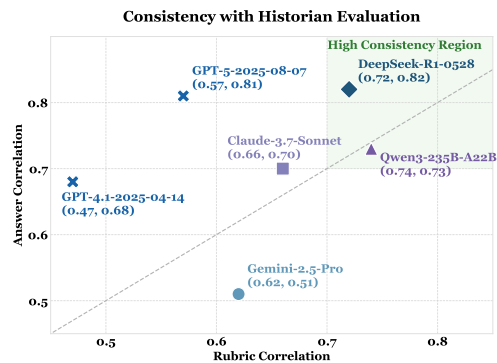


Figure 4: Consistency score of candidate judge models with manually crafted (human) ground truth (answers and rubric).

and collected the outputs of the evaluated LLMs. We asked each candidate judge model and human experts to perform item-by-item 0/1 hit annotation according to the rubric, obtaining fine-grained scores (rubric-level) and an aggregated total score (answer-level). We then computed the Pearson correlation coefficients between each candidate judge model and the experts on both types of scores. As illustrated in Figure 4, DeepSeek-R1 demonstrated the highest average consistency (0.77) with historians at both the answer level and the fine-grained rubric level, and was therefore selected as the judge model.

For scoring, we define a quantitative metric named **Rubric Score (RS)**. The judge model determines the presence (binary 0/1) of each criterion defined in the rubric. The final score is calculated by aggregating the weighted scores of all triggered items (including both bonus and penalty terms) and normalizing by the total potential positive score:

$$RS = \max\left(0, \frac{\sum(I_b \cdot w_b) + \sum(I_p \cdot w_p)}{\sum w_b}\right) \quad (1)$$

where I_b and I_p denote the indicators for bonus and penalty rubric items, respectively ($I \in \{0, 1\}$ denotes the indicator of whether a specific criterion is met), and w represents the pre-defined score of

Model	BL	RG	BS	RS
<i>Closed-Source Models</i>				
Claude-Sonnet-4.5-Thinking	2.53	4.76	71.49	12.99
GPT-5.2	3.50	3.46	71.48	11.07
GPT-5.2-Thinking	4.45	4.58	71.55	14.08
GPT-o3-2025-04-16	7.88	3.65	72.60	14.66
Gemini-3-Pro-Preview	1.94	6.27	73.97	26.71
Gemini-3-Pro-Preview-Thinking	2.35	5.16	73.92	26.73
Qwen3-Max	4.77	6.64	75.01	17.71
<i>Open-Source Models</i>				
Llama-4-Scout-17B-16E	2.59	3.09	72.68	2.72
gpt-oss-120b	1.27	1.78	70.18	10.75
gpt-oss-20b	<i>Fail</i>	<i>Fail</i>	<i>Fail</i>	<i>Fail</i>
Kimi-K2-Thinking	3.62	6.43	73.20	22.79
GLM-4.6-Thinking	2.09	5.11	72.30	24.32
Qwen3-14B-Thinking	1.89	3.61	72.93	11.61
Qwen3-32B-Thinking	2.06	3.97	72.99	13.89
Qwen3-235B-A22B-Thinking	1.08	5.22	72.50	28.14
DeepSeek-V3.2	4.65	6.20	73.52	18.77
DeepSeek-V3.2-Thinking	4.91	6.16	73.41	18.72
DeepSeek-R1-0528	1.93	6.60	73.15	26.87

Table 4: Performance comparison of various LLMs on tasks T1-T3. The reported metrics include average BLEU (BL), ROUGE (RG), BERTScore (BS), and Rubric Score (RS).

that criterion. Negative scores are clipped to 0 to ensure a valid range.

5 Experiments

5.1 Setup

Evaluated Models We conducted extensive experiments on ProHist-Bench, evaluating 7 closed-source LLMs (Claude-Sonnet-4.5-Thinking, GPT-series, Gemini-series and Qwen3-Max) and 11 open-source LLMs (Llama-4-Scout-17B-16E, gpt-oss-series, Kimi-K2-Thinking, GLM-4.6-Thinking, Qwen3-series and DeepSeek-series). We used official APIs or standard deployment setups for all LLMs. To ensure deterministic evaluation, we fixed all hyperparameters and eliminated randomness.

Metric We employ different evaluation metrics tailored to specific task types. We utilize two complementary evaluation approaches: (1) *automatic metrics* including **BLEU (BL)** (Papineni et al., 2002), **ROUGE (RG)** (Lin, 2004), and **BERTScore (BS)** (Zhang et al.) to measure the similarity between LLM outputs and reference answers; and (2) *expert-based evaluation* using **Rubric Score (RS)** to assess the historical research capabilities of LLMs. For T4 task, we rely solely on Rubric Score (RS) for evaluation, as these tasks require nuanced assessment of research quality that cannot be captured by automatic metrics.

5.2 Main results

Core tasks in professional historical research remain extremely difficult for all current LLMs.

As shown in Table 4, even top-performing LLMs such as Gemini-3-Pro and Qwen3-235B achieve RS scores barely approaching 30, while most LLMs score below 15. This pervasive underperformance indicates that existing LLMs still struggle to meet satisfactory professional historical research standards when handling tasks requiring high factual precision, complex logical reasoning, and nuanced understanding of specific historical contexts.

Chinese LLMs demonstrate superior performance in ProHist-Bench.

Qwen3-Max achieves the strongest overall performance with an average BS of 75.01 and secures the top score of 87.08 on T4 task (Figure 5). Notably, open-source LLMs such as DeepSeek-R1-0528 (RS 26.87) follow closely behind, indicating that through large-scale parameters and strategic optimization, the open-source community has achieved capabilities approaching closed-source LLMs. Furthermore, LLMs trained on large-scale Chinese corpus, including GLM-4.6, Kimi-K2, and the DeepSeek series, consistently outperform LLMs like Llama-4-Scout (RS 2.72) and gpt-oss-120b (RS 10.75). This performance gap demonstrates the critical importance of domain-specific pretraining and cultural-linguistic alignment.

5.3 In-Depth Analysis

We evaluated model performance across five key dimensions: prompting strategies, capabilities, difficulty, dynasties, and topics. Here, we highlight the two most critical findings. Detailed analyses of remaining dimensions are provided in Appendix F.

5.3.1 The Impact of Prompting Strategy

To investigate the impact of distinct prompting strategies, we conducted a systematic evaluation on the ProHist-Bench. As shown in Table 5, results demonstrate that Professional Prompting and Role-Playing consistently outperform Chain-of-Thought (CoT). This suggests that for history-literate LLMs, activating expert identity is more effective than enforcing step-by-step reasoning. Conversely, RAG methods, typically effective in knowledge-intensive tasks, underperform in this evaluation due to the severe scarcity of high-quality ancient historical data in existing retrieval corpora. Retrieved document fragments often contain noise or insufficient

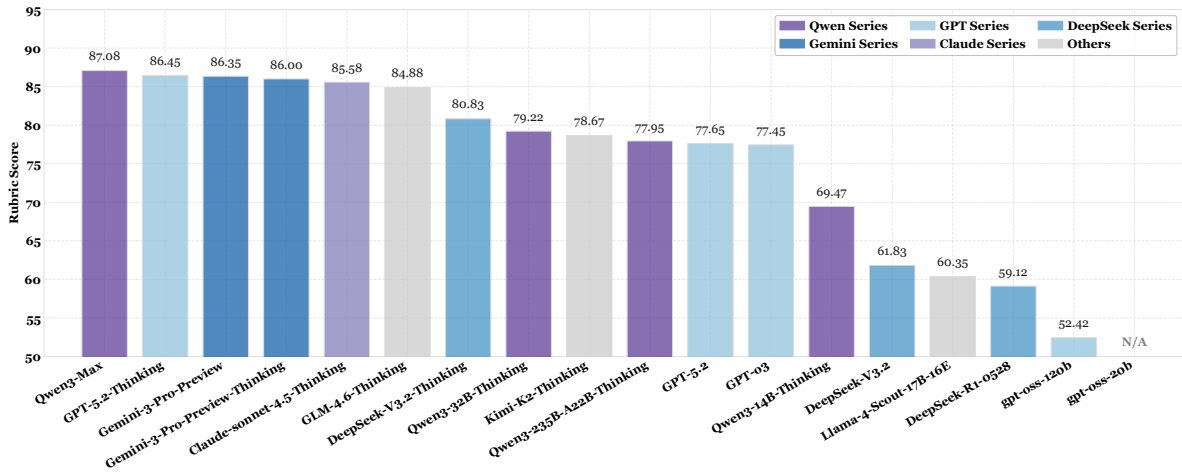


Figure 5: Main results of different LLMs on the T4 task. The table presents the performance of representative closed-source and open-source LLMs. The metrics reported are Rubric Score (RS).

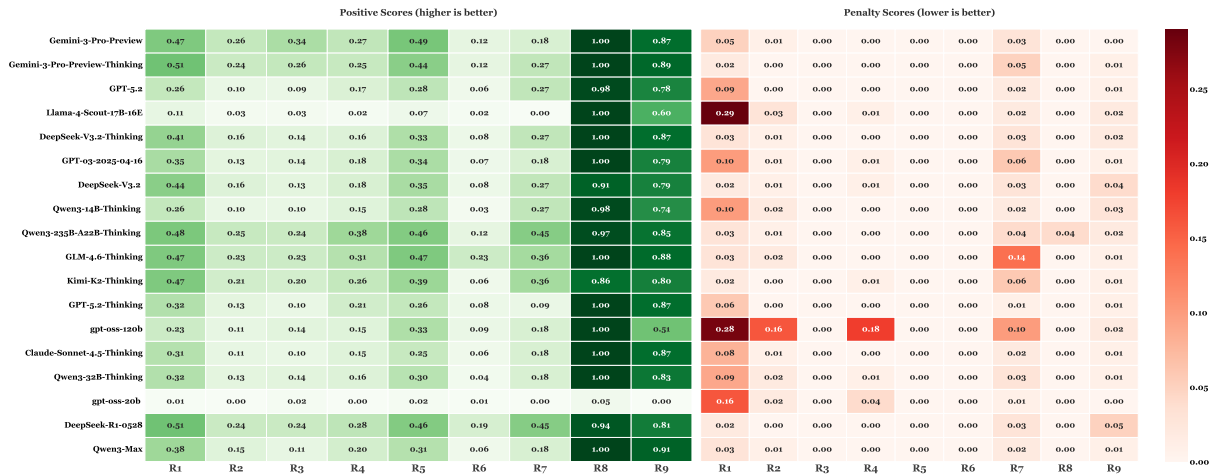


Figure 6: Performance heatmap across nine historical capability dimensions (R1-R9). **Left: Positive Rubric Hit Rate** (darker green indicates better performance). **Right: Penalty Rubric Hit Rate** (darker red indicates higher penalties), quantifying defects such as hallucinations and factual errors.

relevance, disrupting rather than supplementing the model’s reasoning. Detailed RAG experiments and analysis are provided in Appendix E.

5.3.2 Fine-grained Capability Analysis

To comprehensively evaluate the historical research capabilities of LLMs, we analyzed the distribution of Positive and Penalty Rubric Hit Rate (Figure 6), yielding several key observations: **i) Tool-like utility and formalization capabilities are mature.** High positive rubric hit rate in R8 (Classical Writing) demonstrate a mastery of rule-based logic. Furthermore, stable performance in R5 (Comprehensive Evaluation) confirms the models’ ability to generate self-consistent narratives. **ii) Viewpoint Integration (R6) is currently the most prominent shortcoming of existing LLMs.** The R6 hit

rate is extremely low across nearly all models (most are distributed between 0.02 ~ 0.12, with the maximum only at ~ 0.23), indicating that regardless of model capability, there are significant deficiencies in their ability to resolve conflicting historical records. In addition, **iii) capabilities in historical comparison (R3), evidentiary reasoning (R4), and fact organization are also generally weak.** Based on a detailed case study (see Appendix G), we find that the main reasons are twofold: first, most models are unable to distinguish the semantic evolution of the same concept across different historical periods; second, at the levels of reasoning and organization, although models can produce conclusions, they often lack a complete chain of historical evidence and support from a correct timeline, resulting in weak arguments.

Model	Role	Prof.	CoT	RAG
<i>Closed-Source Models</i>				
Claude-Sonnet-4.5-Thinking	19.42	26.01	20.09	15.89
GPT-5.2	15.42	15.73	12.85	14.22
GPT-5.2-Thinking	19.39	19.90	16.96	16.81
GPT-o3	17.09	17.59	14.33	20.29
Gemini-3-Pro-Preview	32.94	30.95	25.08	26.53
Gemini-3-Pro-Preview-Thinking	32.56	31.08	26.26	26.76
Qwen3-Max	24.95	25.05	23.71	20.86
Average	23.11	23.76	19.90	20.19
<i>Open-Source Models</i>				
Llama-4-Scout-17B-16E-Instruct	5.11	4.55	3.88	5.88
gpt-oss-120b	10.61	13.42	10.73	11.71
gpt-oss-20b	0.41	0.77	0.19	2.79
Kimi-K2-Thinking	31.61	36.10	34.10	31.18
GLM-4.6-Thinking	31.73	32.51	27.78	24.10
Qwen3-14B-Thinking	13.12	12.66	11.23	11.92
Qwen3-32B-Thinking	17.23	15.19	15.25	13.33
Qwen3-235B-A22B-Thinking	32.79	34.67	29.72	27.92
DeepSeek-V3.2	28.76	27.72	24.01	27.86
DeepSeek-V3.2-Thinking	29.13	26.68	25.66	26.46
DeepSeek-R1-0528	30.92	26.64	25.96	30.97
Average	21.04	20.99	18.96	19.46

Table 5: Impact of different prompting strategies on LLM performance. We evaluated four strategies: Historian Role-playing (Role), Professional Prompting (Prof.), Chain-of-Thought (CoT), and Retrieval-Augmented Generation (RAG).

5.4 LLMs vs. Human

To further assess LLMs in historical research, we compare SOTA LLMs with two human baselines: *close-book historians*, representing internal knowledge only, and *open-book historians*, representing a reference-augmented upper bound. The human baselines comprise around 20 history professors and PhD students. As shown in Figure 7, SOTA LLMs achieve competitive performance on Historical Reasoning and *Celun* Generation, but still lag far behind open-book historians on precision-critical tasks such as Term Interpretation and Fact QA. This gap suggests that, although LLMs have moved beyond basic historical QA, they have not yet reached the standard required for independent professional verification, and are better viewed as assistive tools rather than autonomous scholars.

6 Conclusion

In this paper, we introduced ProHist-Bench, an expert-curated benchmark anchored in *Chinese Imperial Examination* to rigorously evaluate the historical reasoning capabilities of LLMs. By transcending simple general knowledge QA to focus on professional historical research tasks, our interdisciplinary study reveals a significant proficiency gap between SOTA LLMs and professional his-

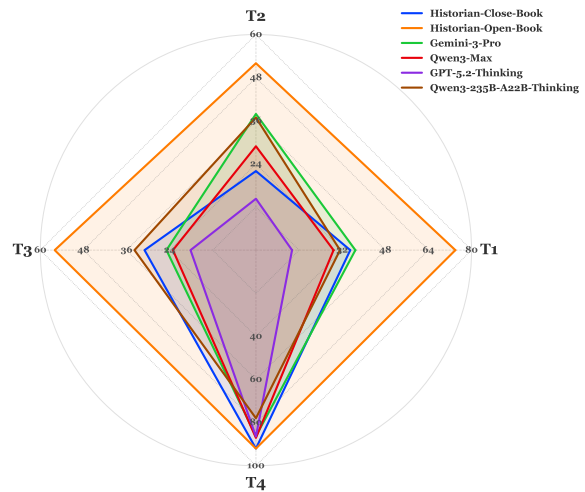


Figure 7: Performance comparison between human experts and LLMs. The Historian-Open-Book represents the expert ceiling with external resource access, while Historian-Close-Book serves as the unassisted human.

torians. Our extensive evaluation highlights that current LLMs still struggle with the complex reasoning required for expert-level analysis.

Limitations

We used a standardized prompt across all 18 models to ensure fairness and comparability. Although some models might perform better with model-specific prompt engineering, our goal was to assess their intrinsic historical reasoning under consistent conditions. Our results provide a robust baseline for performance in the historical domain.

Acknowledgement

This work was supported by Ant Group.

Ethical Considerations

Users of ProHist-Bench should recognize that both historical sources and annotations are inevitably shaped by the limitations of surviving materials and by historiographical interpretation. Accordingly, the dataset should be used with appropriate scholarly caution. ProHist-Bench is intended exclusively for non-commercial academic research, particularly for evaluating LLMs on historically grounded understanding and reasoning. All use must comply with applicable copyright regulations and research ethics standards, and any commercial or harmful use is strictly prohibited.

References

- Anthropic. 2025. [Introducing Claude Sonnet 4.5](#).
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). *CoRR*, abs/2505.08775.
- Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024a. [Wenmind: A comprehensive benchmark for evaluating large language models in chinese classical literature and language arts](#). In *Neural Information Processing Systems*.
- Jiahuan Cao, Yongxin Shi, Dezhi Peng, Yang Liu, and Lianwen Jin. 2024b. [CQ3} bench: A comprehensive classical chinese understanding benchmark for large language models](#). *arXiv preprint arXiv:2405.17732*.
- Y. Z. Chang. 1942. [China and english civil service reform](#). *The American Historical Review*, 47:539.
- Mathieu Chartier, Nabil Dakkoune, Guillaume Bourgeois, and Stephane Jean. 2025. [Hibenchllm: Historical inquiry benchmarking for large language models](#). *Data & Knowledge Engineering*, 156:102383.
- Yuan Chen. 2014. *Shiyuanxue Shixi Ji Qingdai Shixue Kaozhengfa [Practice of Historical Source Study and Textual Research Methods of Qing Dynasty Historiography]*. (in Chinese).
- Zijian Chen, Wenjun Zhang, Guangtao Zhai, and 1 others. 2025. [Obi-bench: Can llms aid in study of ancient script on oracle bones?](#) In *The Thirteenth International Conference on Learning Representations*.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. [Time travel: A comprehensive benchmark to evaluate llms on historical and cultural artifacts](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23627–23641.
- Google DeepMind. 2025. [Gemini 3 Pro: Best for complex tasks and bringing creative concepts to life](#).
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, and 1 others. 2024. [Large language models’ expert-level global history knowledge benchmark \(hist-llm\)](#). *Advances in Neural Information Processing Systems*, 37:32336–32369.
- Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Bo Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. 2024. [Chinese simpleqa: A chinese factuality evaluation for large language models](#). *ArXiv*, abs/2411.07140.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Fanchao Qi, Yao Fu, Maosong Sun, and Junxian He. 2023. [C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models](#). *ArXiv*, abs/2305.08322.
- Zijing Ji, Yuxin Shen, Yin yin Sun, Tian Yu, and Xin Wang. 2021. [C-clue: A benchmark of classical chinese based on a crowdsourcing system for knowledge graph construction](#). In *China Conference on Knowledge Graph and Semantic Computing*.
- Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025. [Flashrag: A modular toolkit for efficient retrieval-augmented generation research](#). In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pages 737–740. ACM.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Tim Baldwin. 2023. [Cmmlu: Measuring massive multi-task language understanding in chinese](#). *ArXiv*, abs/2306.09212.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

- Fangyu Liu, Emanuele Bugliarello, E. Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. [Visually grounded reasoning across languages and cultures](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. [MCS-bench: A comprehensive benchmark for evaluating multimodal large language models in Chinese classical studies](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10435–10492, Vienna, Austria. Association for Computational Linguistics.
- Meta Llama Team. 2025. [The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation](#).
- OpenAI. 2025a. [Introducing GPT-5.2](#).
- OpenAI. 2025b. [Introducing OpenAI o3 and o4-mini](#).
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat et al. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. [gpt-oss-120b gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. [Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2038–2048.
- Qwen Team. 2025. [Qwen3-Max: Just scale it](#).
- Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2023. [M5hisdoc: A large-scale multi-style chinese historical document analysis benchmark](#). *Advances in Neural Information Processing Systems*, 36:78483–78495.
- Yuzhen Shi, Huanghai Liu, Yiran Hu, Gaojie Song, Xinran Xu, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Di Feng, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Yuanyang Qiu, Yueming Qi, Jingwen Zhang, and 11 others. 2026. [Plawbench: A rubric-based benchmark for evaluating llms in real-world legal practice](#). *CoRR*, abs/2601.16669.
- 5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025a. [Glm-4.5: Agentic, reasoning, and coding \(arc\) foundation models](#). *Preprint*, arXiv:2508.06471.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025b. [Kimi k2: Open agentic intelligence](#). *Preprint*, arXiv:2507.20534.
- Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadgign Ademteu, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2024. [All languages matter: Evaluating llms on culturally diverse 100 languages](#). 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19565–19575.
- Yuting Wei, Yuanxing Xu, Xinru Wei, Yangsimin Yangsimin, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. [Ac-eval: Evaluating ancient chinese language understanding in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1600–1617.
- Barbara E. Weinstein. 2005. [History without a cause? grand narratives, world history, and the postcolonial dilemma](#). *International Review of Social History*, 50:71 – 93.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Yanfang Ye, Zheyuan Zhang, Tianyi Ma, Zehong Wang, Yiyang Li, Shifu Hou, Weixiang Sun, Kaiwen Shi, Yijun Ma, Wei Song, and 1 others. 2025. [Llms4all: A systematic review of large language models across academic disciplines](#). *arXiv preprint arXiv:2509.19580*.
- Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2025a. [Can MLLMs understand the deep implication behind Chinese images?](#) In *Proceedings*

of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Yuqing Zhang, Yue Han, Shuanghe Zhu, Haoxiang Wu, Hangqi Li, Shengyu Zhang, Junchi Yan, Zemin Liu, Kun Kuang, Huaiyong Dou, and 1 others. 2025b. Ms-bench: Evaluating llms in ancient manuscript study through a dunhuang case study. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bo Zhou, Qianglong Chen, Xiaomi Zhong, and Yin Zhang. 2023. Wyweb: A nlp evaluation benchmark for classical chinese. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319.

Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. 2025. [Hanfu-bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24627–24649, Suzhou, China. Association for Computational Linguistics.

A Positive Rubric Criteria

To support fine-grained evaluation, we design nine rubric dimensions covering different aspects of historical understanding, reasoning, and generation. Detailed definitions are provided below.

R1. Concept Definition. Assesses whether the model can provide concise conceptual definitions of historical phenomena before detailed elaboration, ensuring the definition captures core elements while maintaining historical accuracy. (2 points)

R2. Fact Organization. As the most fundamental capability in historical research, this evaluates whether the model can clearly and comprehensively narrate the basic processes of events, institutions, phenomena, or intellectual developments from emergence through evolution to dissolution. (3 points)

R3. Historical Comparison. Building upon fact organization, this assesses the ability to identify historical evolution and cultural variations across different dynasties anchored in the *Keju*, demonstrating cross-temporal and cross-spatial comparative capabilities. (3 points)

R4. Evidentiary Reasoning. Evidentiary Reasoning assesses the fundamental principle of arguments grounded in historical evidence by evaluating whether the model can derive reliable and academically consensual conclusions after organizing facts. This dimension ensures that all arguments are substantiated by credible historical sources rather than hallucinations and requires the model to cite concrete historical examples to support its viewpoints. (4 points)

R5. Comprehensive Evaluation. Examines whether the model can provide holistic evaluations and interpretations of the significance of historical events, figures, and institutions—a fundamental capability. (1 point)

R6. Viewpoint Integration. Evaluates whether the model can incorporate cutting-edge scholarship and research trends, present mainstream academic perspectives on historical phenomena, and establish clear connections among different scholars' viewpoints. (5 points)

R7. Academic Expression. Assesses the rigor and professionalism of the model's academic writing. (5 points)

R8. Classical Writing. Evaluates structural completeness based on essential elements of the eight-legged essay format, including opening (破题), elaboration (承题), transition (起讲), and thesis (入题). (3 points)

R9. Temporal Reframing. Evaluates the ability of LLM to reason within a specific historical context. Specifically, the rubric allocates 4 points to whether the writing style accords with the preferences of a given dynasty and 5 points to whether the essay meets basic compositional requirements. In the *Celun* Generation task, the LLM is positioned as a candidate in the imperial examination of a particular dynasty, and the evaluation therefore also considers its grasp of the institutional context, literary conventions, and historical taboos (避讳词) of the period. To preserve historical fidelity, we impose substantial penalties. Responses that

depart from period style incur a 40 point deduction, whereas responses that violate examination norms incur a 60 point deduction. For example, the use of taboo words such as an emperor’s name could historically result in disqualification, and the 60 point penalty is intended to reflect the seriousness of such errors.

B General Penalty Rubric Criteria

General penalty rubric criteria includes:

- (1) **Inappropriate Academic Formulation:** Instances of mechanically applying social science theories or making inappropriate analogies. *{-3 points}*
- (2) **Fabrication of References or Materials:** Acts of fabricating cited references or inventing historical sources/materials. *{-5 points}*
- (3) **Missing Citations:** Failure to provide standard citations when quoting Classical Chinese texts. *{-3 points}*
- (4) **Chronological Conversion Errors:** Errors in converting ancient Chinese era names to Common Era (AD) years. *{-3 points}*
- (5) **Core Concept Errors:** A lack of understanding or complete misinterpretation of specific historical terms. *{-5 points}*
- (6) **Non-Academic Language Style:** The content is overly colloquial and lacks the rigor expected of academic writing. *{-1 points}*

The criteria listed above apply to the entire evaluation process and are organized along two dimensions: i) Factuality Dimension (Items 1–5): Designed to assess whether the model can accurately reflect historical records at the level of historical fact. Items 1, 2, and 3 focus on detecting *hallucinations*—instances where the model generates content that appears plausible and self-consistent but lacks a factual basis. Items 4 and 5 focus on detecting the model’s knowledge accuracy, checking for factual errors during the generation of specific knowledge points. ii) Expression Dimension (Item 6): Designed to assess the model’s command of academic language. It requires that the generated content adheres to the academic norms of the humanities and social sciences, avoiding overly colloquial, entertaining, or casual unprofessional expressions.

C Dataset Scope and Topic Taxonomy

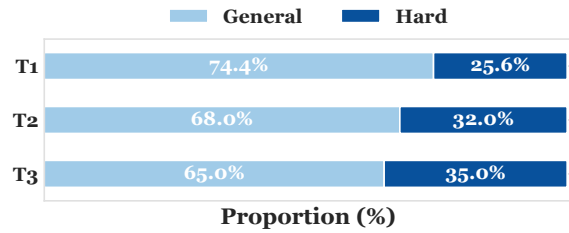


Figure 8: Distribution of difficulty levels (General vs. Hard) across different task categories.

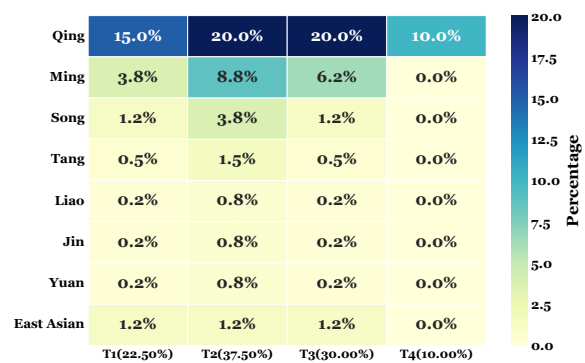


Figure 9: Cross-distribution of dynasties and task categories. Darker colors indicate higher proportions. “East Asia” represents historical knowledge of East Asian countries other than China.

We present the statistical data of ProHist-Bench in Figures 8, 9 and 3. ProHist-Bench consists of 400 expert-curated challenging historical research questions across four core tasks: Term Interpretation, Fact QA, Historical Reasoning, and *Celun* Generation.

C.1 Difficulty Distribution

Difficulty distribution (Figure 8): Historical Reasoning is the most challenging, with 35.0% hard samples, compared to only 25.6% for Term Interpretation.

C.2 Task and Dynasty Distribution

Task distribution (Figure 9): Fact QA dominates at 37.50%, while *Celun* Generation is minimal at 10.00%. Historical period distribution (Figure 9): The Qing Dynasty predominates, particularly in Term Interpretation and Historical Reasoning (both 20.0%), due to its proximity to modern times and well-preserved documentation.

C.3 Topic Taxonomy and Distribution

To ensure a systematic and comprehensive evaluation of LLMs, we constructed a fine-grained knowledge taxonomy, as illustrated in Figures 10. This hierarchical framework decomposes the complex historical domain into **13 Main Topics** and their corresponding **31 Sub-topics**.

Unlike traditional benchmarks that focus primarily on factual recall, our taxonomy spans a broad spectrum of historiographical dimensions. It ranges from fundamental institutional procedures (e.g., *Keju Registration & Process, Offices and Roles*) to complex socio-political interactions (e.g., *The Interplay between Keju and Socio-economy, Social Networks of Literati*). Furthermore, it incorporates high-level historical analysis, such as *Keju Reform & Abolition* and *Keju's Value in Human Civilization*. This structured approach serves as the foundational guide for data collection, ensuring that the benchmark not only tests the model's knowledge retention but also its capability to handle diverse historical contexts and conduct deep reasoning across different granularity levels.

D Extended Discussion on Prompting Strategies

As illustrated in Table 6, we analyzed how different prompting methods—including Role, Prof., CoT, and RAG—affect the capabilities of various LLMs. Table 6 details performance metrics across Term Interpretation, Fact QA, and Historical Reasoning, highlighting significant variations between closed-source and open-source LLMs under different prompting conditions. The specific prompt templates utilized in these experiments are provided in Figure 11 and Figure 12.

E Detailed Evaluation of RAG

Implementation of RAG. We construct the retrieval corpus by cleaning, segmenting, and vectorizing the latest Chinese Wikipedia dump. Our RAG system is implemented with the FlashRAG framework (Jin et al., 2025), using bge-large-zh-v1.5* as the embedding model.

In this section, we examine how varying the number of retrieved documents (k) affects model performance. As shown in Table 7, increasing k does

*<https://huggingface.co/BAAI/bge-large-zh-v1.5>

not lead to consistent gains. Instead, most LLMs exhibit substantial sensitivity to retrieval noise. For example, for Gemini-3-Pro-Preview, the average score decreases from 27.35 to 25.51 as k increases from 10 to 100. This suggests that, for models with relatively strong internal knowledge, retrieving too many external documents (e.g., $k = 50$ or 100) may introduce irrelevant information that interferes with otherwise precise reasoning.

F Additional Experimental Analysis and Results

F.1 Difficulty-wise Performance Analysis

As illustrated in Figure 13, the transition from General to Hard tasks precipitates a universal performance decline, with the average score plummeting from 20.89 to 8.39. This $> 50\%$ drop confirms that while current LLMs possess broad historical knowledge, they lack the deep inferential capabilities required for complex historiography. Notably, models incorporating chain-of-thought mechanisms—such as Gemini-3-Pro-Preview-Thinking (15.62) and Qwen3-235B-A22B-Thinking—demonstrate superior resilience compared to standard models, indicating that explicit reasoning steps are essential for navigating complex historical logic.

F.2 Topic-wise Performance Analysis

As detailed in Table 8, model performance exhibits sharp variance across historical sub-domains, revealing a distinct dichotomy between static knowledge and complex reasoning. Specifically, while models consistently achieve higher scores in factual tasks like *Keju Literature* and *Ancient Education*, their proficiency drops notably in reasoning-intensive domains such as *Keju Reform* and *Ancient Geo & Transport*, which demand the synthesis of spatial and temporal changes. This analysis also highlights a significant capability gap: top-tier closed-source models (e.g., Gemini-3-Pro) maintain robustness across diverse topics, whereas open-source models struggle with complex contexts—exemplified by near-zero scores in *Ancient Social Networks* due to their inability to capture implicit interpersonal relations.

F.3 Dynasty-Level Performance Analysis

Figure 14 reveals a marked imbalance in model proficiency across historical eras, correlating strongly

Main Topic	Sub-topic	Description
科举相关制度的定义与演变 Keju Definition & Evolution	科举制度沿革	科举制度的起源、发展、历代变化等
	科举考试类别	科举制度中的考试类型
	科举考试科目	科举考试科目的起源、发展、历代变化等
官职体系与科举关联 Offices and Roles in Keju	科举功名类别	对科举中不同出身的介绍，包括基本概念、权利义务以及相关的规章制度。
	科举职官制	科举制中的职官体系，包括考试官、阅卷官、读卷官等。
报名方式与考试流程 Keju Registration & Process	应试资格	报名资格、准入条件
	考试层级	科举考试的层级指的是童试、乡试、会试/（发）解试/省试、殿试四个层级，根据不同朝代，考试层级会发生变化。其他科目考试并不算考试层级，属于考试科目。
	考试内容与题型	经义、策论、诗赋、八股结构、考试题目等
	考务管理	考官选派、试卷管理、贡院安排等和考场、考试相关的内容、评分标准
	防弊制度与弊案	防止舞弊的措施及历史案例如作弊
录取机制与授官路径 Admission and Official Appointment	录取机制	录取标准、方式、名额
	科举与授职关系	科举录取后与授官的关系
	中举仪式	传胪、琼林宴、游街等录取庆典
	落第	落第举子的社会状况与经历以及落第的相关政策
科举与社会、经济的互动 The Interplay between Keju and socioeconomy	科举费用	报名、考试的全部费用及相关物品
	社会经济支持	宾兴、学田、宗族义庄等社会协助
	科举与社会文化生活	相关习俗、出版印刷、婚姻、文化现象等
同年关系与社交网络 Social Networks of Literati	座主门生关系与官场结党	师徒、同门、党派关系
	学术传承与考试内容变革	学术流派、考试内容的变化与学术发展
考试物质与生活场域 Examination Tools and Space	贡院	考场环境设置与发展
	科场供应	考试期间饮食、文具、住宿等物质生活
学校与教育 Ancient Education	官学	国子监、官办学堂、地方府州县学、晚清学堂等
	私学	非官方兴办的教学机构，如书院、义学等民间教育机构。
地理与交通 Ancient Geo & Transport	科举群体的地理分布与差异	区域间科举差异、各省份举子状况
	区域制度差异	不同地区相关科举政策与制度
科举制度改革与废除 Keju Reform & Abolition	历史上重要的科举制度改革	历史上会对科举制度进行改革，要区分清制度改革和制度变化。界定改革和变化的关键在于：改革是系统的、集中的，而变化是非系统的、临时的、细碎的。其次一次重要的科举改革会对当时的社会、政治、文化产生深刻的影响，形成显著的社会效应。第三，改革会引发古今社会的广泛议论和关注。
	清代废科举	清末废除科举相关问题
科举与学术思想 Ancient Academic Thoughts	学术思潮变迁与科举的双向影响	科举制度与学术思想之间的相互关系
科举文献 Keju Literature	科举文献	与科举相关的档案、书籍、文献等
科举制在人类文明史上的地位 Keju's Value in Human Civilization	亚洲国家对科举制的借鉴	主要包括亚洲国家对科举制的移植、模仿、学习、本土化。
	科举制的国际化影响	主要包括科举制在世界的地位，科举制对其他国家制度建设、社会文化等方面产生的影响。

Figure 10: The hierarchical taxonomy of the topic framework for ProHist-Bench. This framework categorizes the dataset into 13 Main Topics and their corresponding Sub-topics, covering diverse dimensions ranging from institutional procedures and official roles to socio-economic interactions and global influences.

with textual data availability. While top-tier LLMs (e.g., Qwen3-235B-Thinking, Gemini-3-Pro) demonstrate robustness in mainstream dynasties such as Tang, Song, and Ming—achieving peak RS exceeding 30.0—performance universally degrades in low-resource contexts like the Liao and

Jin dynasties, where scores often drop significantly. This result highlight that current LLMs struggle to transfer reasoning capabilities to historical periods with sparser textual records.

Prompt Templates of Role Play and Professional Prompting

Role Play:

System: 你是一位资深的中国历史学家与研究者，专注于严谨的史学研究及解读。你的回答应基于史料与历史学方法，保持信息准确、逻辑清晰和学术性。在回答时：

- 使用精炼的学术语言，避免娱乐化或泛泛而谈的通俗解释。
- 明确时间、地点、人物、背景、史实细节。
- 指出存在的学术争议和不同观点的差异。
- 避免臆测或虚构史实，确保回答可被史料验证。

目标是：让回答具备学术讨论价值，而不仅是简单的知识介绍。

请回答下面的问题：

User: {Question}

System: You are a senior historian and researcher specializing in Chinese history, dedicated to rigorous historiographical research and interpretation. Your responses must be grounded in historical materials and methodologies, ensuring factual accuracy, logical clarity, and academic rigor.

When responding:

1. Use concise, scholarly language; avoid sensationalism or superficial, generalized explanations. 2. Explicitly specify dates, locations, key figures, historical contexts, and factual details.
3. Identify existing academic controversies and the nuances of diverging scholarly perspectives.
4. Avoid conjecture or fabrication; ensure all content is verifiable through historical sources.

The goal is to ensure your responses possess value for academic discourse, rather than serving merely as simple informational introductions.

Please answer the following question:

User: {Question}

Professional Prompting:

System: 你是一位熟练使用 **5W1H** 历史分析法的中国历史研究者。你的任务是基于用户的提问，以全面的角度拆解事件或问题，并在输出中体现以下分析维度：

1. **What:** 事件的定义、特征、核心构成要素、史实描述
2. **Why:** 直接与根本原因、背景、动机、逻辑关系、学术视角
3. **When:** 时间定位、阶段划分
4. **Who:** 相关的历史主体、身份来源、动机、权力关系
5. **Where:** 地点与区域差异影响
6. **How:** 事件的过程、运作方式、机制分析

在回答中可根据问题选择相关维度，但必须保证分析充分且结构化。

回答要求：如某一维度存在史学争议，需列出不同观点并说明依据。信息需符合史料记载，并保持逻辑严密。输出应按维度分段，便于阅读与评估。

请回答下面的问题：

User: {Question}

System: You are a researcher of Chinese history skilled in employing the **5W1H Historical Analysis Method**. Your task is to comprehensively deconstruct events or issues based on user inquiries, incorporating the following analytical dimensions into your response:

1. **What:** Definition of the event, characteristics, core constituent elements, and description of historical facts
2. **Why:** Direct and root causes, background, motivations, logical relationships, and academic perspectives
3. **When:** Temporal localization and periodization
4. **Who:** Relevant historical actors, origins of identity, motivations, and power relations
5. **Where:** Geographic location and the impact of regional differences
6. **How:** The process of the event, modes of operation, and mechanism analysis

You may select relevant dimensions based on the specific question, but you must ensure the analysis is thorough and structured.

Response Requirements:

- If a specific dimension involves historiographical controversy, you must list divergent viewpoints and explain their evidentiary basis.
- Information must align with historical records and maintain rigorous logic.
- The output should be segmented by dimension to facilitate reading and evaluation.

Please answer the following question:

User: {Question}

Figure 11: Prompt templates of Role Play and Professional Prompting.

Prompt Templates of CoT and RAG

CoT:

System: 你是一位具有史学研究训练的中国历史学者，你在回答问题时必须使用链式推理（**Chain of Thought**）方法，逐步分析并得出结论。回答要求：

- 每一步推理必须明确且条理清晰
- 所有历史信息需基于可信史料
- 如果有不同推断路径，应标明其可能性与依据
- 保持严谨性与透明性，让读者能够清楚看到你的思考链条

让我们一步步思考，回答下面的问题：

User: {Question}

System: You are a scholar of Chinese history with training in historiographical research. When answering questions, you must use the **Chain of Thought** method to analyze step-by-step and reach a conclusion. Response Requirements:

- Each step of reasoning must be explicit and clearly organized.
- All historical information must be based on credible historical sources.
- If there are divergent inferential paths, indicate their plausibility and evidentiary basis.
- Maintain rigor and transparency, allowing the reader to clearly see your chain of reasoning.

Let's think step by step to answer the following question:

User: {Question}

RAG:

System: 你是一位精通历史的研究助手，擅长结合自身知识和外部资料为历史学家提供准确、权威的回答。请基于你的知识储备回答用户的问题，生成准确、自然的类似人类的回答。以下是一些可能相关的参考文档。请注意：

- 这些文档可能与问题相关，也可能不完全相关
- 如果文档内容能够帮助回答问题，请适当参考
- 如果文档与问题关联不大，请主要依靠你自身的知识回答
- 优先保证回答的准确性和相关性。

请回答下面的问题：

User: {Question}

System: You are a research assistant proficient in history, skilled in combining your own knowledge with external resources to provide accurate and authoritative answers for historians. Please answer the user's question based on your knowledge base, generating an accurate, natural, and human-like response. Below are some potentially relevant reference documents. Please note:

- These documents may be relevant to the question, or they may not be entirely relevant.
- If the content of the documents helps answer the question, please reference them appropriately.
- If the documents are not significantly related to the question, please rely primarily on your own knowledge to answer.
- Prioritize the accuracy and relevance of your response.

Please answer the following question:

User: Question

Figure 12: Prompt templates of CoT and RAG.

F.4 Detailed Results of Different LLMs on the T1-T3 Tasks

Detailed results of different LLMs on the T1-T3 tasks are reported in this section. Table 9 presents the performance of representative closed-source and open-source LLMs

F.5 Detailed Comparison between Models and Human Performance

As presented in Table 10, top-tier LLMs (e.g., Gemini-3-Pro-Preview, Qwen3-235B-Thinking) effectively surpass the *Closed-book Historian* baseline (Avg ~34.0 vs. 29.33), demonstrating superior large-scale knowledge retention. However, a substantial gap persists when compared to the *Open-book Historian* (60.67), highlighting that while models excel at static retrieval, they significantly lag behind human experts in complex evidence syn-

Model	Term Interpretation				Fact QA				Historical Reasoning				Average			
	Role	Prof.	CoT	RAG	Role	Prof.	CoT	RAG	Role	Prof.	CoT	RAG	Role	Prof.	CoT	RAG
<i>Closed-Source Models</i>																
Claude-Sonnet-4.5-Thinking	18.65	26.53	20.50	15.77	17.37	27.34	21.23	20.28	23.43	23.95	18.37	13.18	19.42	26.01	20.09	15.89
GPT-5.2	16.66	15.01	11.78	14.15	13.76	15.95	12.90	22.11	15.58	15.99	13.58	9.21	15.42	15.73	12.85	14.22
GPT-5.2-Thinking	22.97	19.01	16.31	17.04	15.73	21.23	17.61	21.89	18.29	18.90	16.65	13.32	19.39	19.90	16.96	16.81
GPT-o3	18.03	19.23	16.13	17.12	13.27	18.84	15.39	30.11	20.62	14.80	11.67	16.93	17.09	17.59	14.33	20.29
Gemini-3-Pro-Preview	31.95	37.64	31.20	27.88	29.19	30.83	24.51	28.44	39.59	26.08	21.21	24.04	32.94	30.95	25.08	26.53
Gemini-3-Pro-Preview-Thinking	32.93	34.78	32.88	24.85	27.74	30.61	25.59	27.06	38.37	28.89	22.13	28.36	32.56	31.08	26.26	26.76
Qwen3-Max	24.70	29.27	26.90	22.46	20.80	24.39	24.86	21.78	30.89	22.73	19.88	18.79	24.95	25.05	23.71	20.86
Average	23.70	25.93	22.24	19.90	19.69	24.17	20.30	24.52	26.68	21.62	17.64	17.69	23.11	23.76	19.90	20.19
<i>Open-Source Models</i>																
Llama-4-Scout-17B-16E-Instruct	4.74	6.87	5.40	5.46	3.06	4.33	3.69	8.72	8.44	3.09	2.98	4.43	5.11	4.55	3.88	5.88
gpt-oss-120b	12.75	10.58	8.12	10.08	8.56	16.79	13.20	16.89	9.80	11.35	9.61	9.89	10.61	13.42	10.73	11.71
gpt-oss-20b	0.12	1.00	0.24	1.65	0.19	0.91	0.07	6.17	1.20	0.42	0.31	1.68	0.41	0.77	0.19	2.79
Kimi-K2-Thinking	30.99	42.68	40.21	27.42	27.96	35.37	32.86	36.94	37.50	32.07	31.07	30.96	31.61	36.10	34.10	31.18
GLM-4.6-Thinking	32.67	35.73	30.04	20.85	28.10	33.95	27.99	24.78	35.01	28.28	25.83	26.68	31.73	32.51	27.78	24.10
Qwen3-14B-Thinking	13.01	15.89	12.62	11.62	11.07	11.31	10.49	17.06	16.03	11.93	11.09	8.89	13.12	12.66	11.23	11.92
Qwen3-32B-Thinking	17.63	18.09	17.14	14.62	13.68	15.41	16.13	16.11	21.30	12.74	12.73	10.36	17.23	15.19	15.25	13.33
Qwen3-235B-A22B-Thinking	33.29	38.93	34.79	28.92	29.23	36.12	30.75	31.11	36.69	29.67	24.63	24.93	32.79	34.67	29.72	27.92
DeepSeek-V3.2	28.25	31.32	28.69	26.31	24.22	28.37	25.29	29.06	35.69	24.20	18.89	28.54	28.76	27.72	24.01	27.86
DeepSeek-V3.2-Thinking	29.41	33.66	31.53	23.69	24.02	25.25	25.46	30.78	35.46	23.23	21.50	26.25	29.13	26.68	25.66	26.46
DeepSeek-R1-0528	29.39	27.50	27.89	30.00	26.96	27.45	27.08	34.28	38.77	25.00	23.11	29.75	30.92	26.64	25.96	30.97
Average	21.11	23.84	21.52	18.24	17.91	21.39	19.36	22.90	25.08	18.36	16.52	18.40	21.04	20.99	18.96	19.46

Table 6: Impact of different prompting strategies on model performance. The table compares the effectiveness of four prompting methods: **Historian Role-playing (Role)**, **Professional Prompting (Prof.)**, **Chain-of-Thought (CoT)**, and **Retrieval-Augmented Generation (RAG)** across Term Interpretation, Fact QA, and Historical Reasoning tasks.

thesis and research-oriented reasoning.

G Case Study

We conducted detailed case studies on Tasks T1-T4, as shown in Figures 15, 16, 17, 18, 19, and 20. By analyzing the T1 to T4 cases, we can draw the following conclusions about LLMs’ capabilities in historical research: (1) **Lack of fine-grained knowledge:** In T1 and T2 task, the model often provides correct “general definitions”, but when faced with rubric-required specific details (e.g., the names of specific punishments, or exact time points in the evolution of certain official titles), it is prone to omissions or hallucinations. (2) **Temporal confusion:** In the T3 (historical reasoning) task, the model tends to confuse institutional differences across dynasties (e.g., Tang–Song vs. Ming–Qing). For example, when explaining the “metropolitan examination” (Huishi, 会试), some models fail to accurately distinguish that the term became standardized in the Ming–Qing, and instead make a loose analogy by mapping it to the Tang–Song “provincial examination” (Shengshi, 省试). (3) **Differences in hard-constraint adherence:** T4 (*Celun* Generation) shows the greatest variation among models. This is the most difficult task because it re-

quires the model to satisfy both format constraints (the baguwen structure) and negative constraints (name taboos). Some LLMs (e.g., Gemini-3-Pro) perform excellently on both taboos and structure, while others (e.g., Qwen3-235B) received a score of 0 for failing to follow the taboo or format requirements.

H Specific Source of ProHist-Bench

Ancient Chinese Text, including Ancient books, local gazetteers, original archival compilations, and eyewitness accounts:

- (清) 奎润等纂修, 李兵、袁建辉点校: 《钦定科场条例》, 长沙: 岳麓书社, 2019年.
- (清) 素尔讷等纂修, 霍有明、郭海文校注: 《钦定学政全书校注》, 武汉: 武汉大学出版社, 2009年.
- 《浙江通志》编纂委员会: 《浙江通志》第76卷《教育志一》, 杭州: 浙江人民出版社, 2019年.
- 顾廷龙主编: 《清代朱卷集成》, 台北: 成文出版社, 1992年.

Model	RAG ($k = 10$)				RAG ($k = 20$)			
	Concept Interpretation	Fact Exposition	Historical Reasoning	Average	Concept Interpretation	Fact Exposition	Historical Reasoning	Average
<i>Closed-Source Models</i>								
Claude-sonnet-4.5-20250929-Thinking	11.86	15.33	16.53	14.25	12.65	21.82	16.50	16.10
GPT-5.2	14.41	23.46	19.05	18.65	11.81	22.24	18.04	16.35
GPT-5.2-Thinking	16.10	21.54	21.16	19.25	16.39	26.71	19.75	19.94
GPT-o3-2025-04-16	20.28	19.79	24.58	21.25	20.84	27.76	21.42	22.67
Gemini-3-Pro-Preview	25.17	28.83	28.79	27.35	26.48	24.59	26.50	26.04
Gemini-3-Pro-Preview-Thinking	22.62	28.88	32.11	27.21	24.97	24.82	24.83	24.89
Qwen3-Max	17.24	24.17	24.95	21.58	17.94	26.00	24.17	21.92
<i>Open-Source Models</i>								
Llama-4-Scout-17B-16E-Instruct	4.62	6.54	3.74	5.03	5.00	7.12	7.96	6.49
gpt-oss-120b-mxfp4-high	9.97	12.08	13.42	11.58	11.61	14.35	11.42	12.19
gpt-oss-20b-mxfp4-high	1.72	2.38	0.37	1.58	1.19	2.88	6.96	3.51
Kimi-K2-Thinking	23.62	25.29	24.47	24.40	22.19	28.82	25.79	24.96
GLM-4.6-Thinking	13.69	22.21	20.42	18.31	17.10	22.94	15.63	17.99
Qwen3-14B-Thinking	8.90	14.67	12.00	11.64	9.16	16.82	11.79	11.85
Qwen3-32B-Thinking	11.03	13.17	14.47	12.65	12.81	11.29	14.29	12.94
Qwen3-235B-A22B-Thinking	17.93	18.29	22.16	19.17	16.74	20.94	15.25	17.24
DeepSeek-V3.2	21.07	23.92	22.89	22.50	23.55	24.65	24.75	24.21
DeepSeek-V3.2-Thinking	21.55	24.54	28.53	24.39	17.81	30.24	23.96	22.79
DeepSeek-R1-0528	25.24	24.58	26.00	25.22	20.13	32.35	29.13	26.01
Model	RAG ($k = 50$)				RAG ($k = 100$)			
	Concept Interpretation	Fact Exposition	Historical Reasoning	Average	Concept Interpretation	Fact Exposition	Historical Reasoning	Average
<i>Closed-Source Models</i>								
Claude-sonnet-4.5-20250929-Thinking	20.72	12.13	11.77	14.13	16.17	16.11	7.71	12.51
GPT-5.2	24.17	15.13	16.58	18.01	17.96	19.67	13.29	16.38
GPT-5.2-Thinking	27.11	19.83	15.52	19.79	19.52	24.17	18.16	20.10
GPT-o3-2025-04-16	28.17	22.17	18.03	21.89	22.09	32.39	23.03	25.07
Gemini-3-Pro-Preview	27.06	27.87	22.81	25.49	26.09	28.22	23.52	25.51
Gemini-3-Pro-Preview-Thinking	27.28	23.04	20.32	22.93	26.22	24.94	20.94	23.63
Qwen3-Max	26.28	17.65	18.26	20.07	19.65	26.28	19.26	21.14
<i>Open-Source Models</i>								
Llama-4-Scout-17B-16E-Instruct	9.11	4.65	7.45	6.97	7.39	7.44	6.26	6.92
gpt-oss-120b-mxfp4-high	19.56	11.57	5.58	10.99	12.83	9.67	7.74	9.85
gpt-oss-20b-mxfp4-high	3.78	0.61	1.97	1.99	0.61	1.39	1.58	1.22
Kimi-K2-Thinking	31.67	21.30	24.19	25.14	16.57	26.28	19.77	20.38
GLM-4.6-Thinking	19.00	16.09	21.42	19.11	17.35	19.17	17.00	17.65
Qwen3-14B-Thinking	16.56	7.35	11.19	11.31	7.70	18.22	8.84	10.82
Qwen3-32B-Thinking	18.89	11.78	9.52	12.58	12.57	17.94	10.58	13.06
Qwen3-235B-A22B-Thinking	18.78	13.70	15.74	15.85	16.78	31.89	14.74	19.68
DeepSeek-V3.2	25.44	21.39	20.19	21.89	22.22	24.72	20.19	21.97
DeepSeek-V3.2-Thinking	33.78	18.91	20.16	23.17	22.26	24.00	22.77	22.92
DeepSeek-R1-0528	25.83	19.65	25.16	23.57	18.17	23.94	23.58	21.94

Table 7: Performance comparison of LLMs with varying numbers of retrieved documents (k) provided as context. Here, $k \in \{10, 20, 50, 100\}$ denotes the specific number of top-ranked documents retrieved and concatenated into the input prompt for each model.

• 李德龙、董：《未刊清代朱卷集成》，北京：学苑出版社，2019年。

• 商衍鏊：《清代科举考试述录》，北京：故宫出版社，2012年。

Authoritative Scholarly Monographs includes:

• 陈维昭、张文达：《张文达藏稀见清代科举文献汇编》，桂林：广西大学出版社，2022年。

• 齐如山：《中国的科名》，杭州：浙江古籍出版社，2020年。

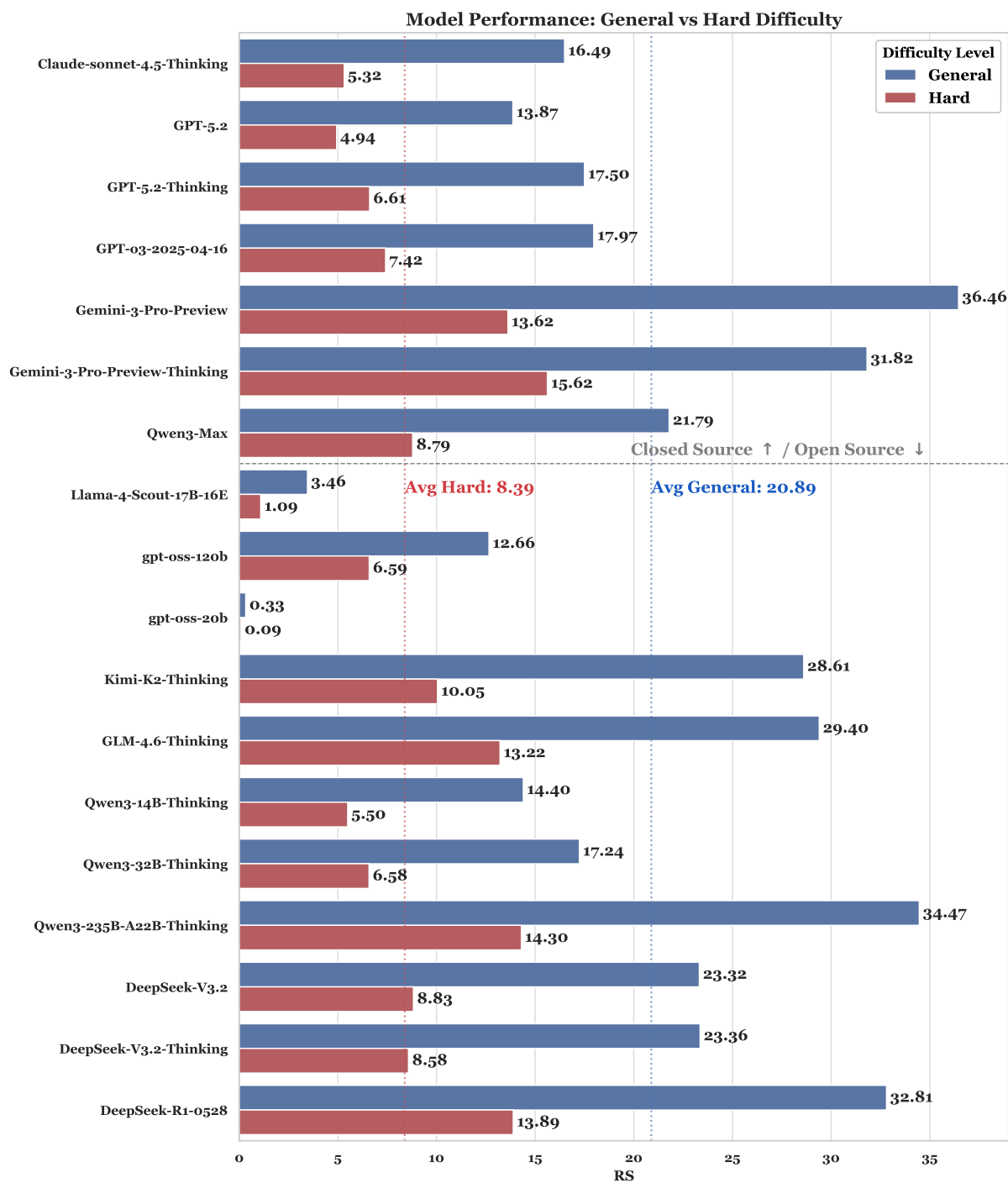


Figure 13: Performance gap between General and Hard difficulty levels. The sharp decline in average scores (20.89 to 8.39) confirms the benchmark's rigor in distinguishing deep historical reasoning from basic factual recall.

- 何怀宏：《选举社会及其终结——秦汉至晚清历史的一种社会学阐释》，北京：生活·读书·新知三联书店，1998年。
- 李兵：《书院与科举关系研究》，武汉：华中师范大学出版社，2005年。
- 李兵、刘海峰：《科举不只是考试》，上海：上海教育出版社，2018年。
- 李华瑞主编：《“唐宋变革论”的由来与发展》，天津：天津古籍出版社，2010年。
- 冯建民：《清代科举与经学关系研究》，武汉：华中师范大学出版社，2016年。
- 贺晓燕：《清代科举落第制度研究》，广州：广东人民出版社，2022年。
- 蒋维明：《李调元传》，北京：天地出版

Model	Ancient Social Net	Ancient Geo & Transport	Ancient Education	Ancient Official System	Admission & Career Path	Keju Registration & Process	Ancient Academic Thoughts
<i>Closed-Source Models</i>							
Claude-sonnet-4.5-20250929-Thinking	24.00	11.17	13.75	15.70	15.23	12.34	19.00
GPT-5.2	15.00	15.28	11.30	12.07	12.88	12.07	16.75
GPT-5.2-Thinking	27.80	10.11	15.00	17.26	18.85	14.18	23.00
GPT-o3-2025-04-16	22.80	18.22	15.50	18.61	17.07	13.81	21.88
Gemini-3-Pro-Preview	36.20	27.33	23.02	31.22	27.48	26.82	39.88
Gemini-3-Pro-Preview-Thinking	25.80	29.83	26.48	29.43	28.45	25.62	36.63
Qwen3-Max	27.40	22.72	16.14	20.54	20.93	18.12	23.50
<i>Open-Source Models</i>							
Llama-4-Scout-17B-16E-Instruct	3.20	2.44	2.36	2.72	2.48	2.78	3.75
gpt-oss-120b	15.80	13.50	7.86	10.59	10.90	10.78	22.38
gpt-oss-20b	0.00	1.22	0.20	0.11	0.35	0.20	0.00
Kimi-K2-Thinking	35.40	27.11	25.09	24.52	26.37	22.00	40.25
GLM-4.6-Thinking	41.20	28.72	25.23	30.57	23.07	24.40	32.25
Qwen3-14B-Thinking	20.60	8.94	13.29	14.52	13.78	10.66	14.50
Qwen3-32B-Thinking	17.20	10.94	13.70	19.87	16.47	14.66	22.50
Qwen3-235B-A22B-Thinking	52.00	30.33	30.84	32.50	29.73	26.20	40.88
DeepSeek-V3.2	23.60	18.28	19.23	21.72	21.58	18.54	28.00
DeepSeek-V3.2-Thinking	22.60	22.56	18.34	23.35	23.92	18.82	22.50
DeepSeek-R1-0528	33.80	30.44	27.13	32.52	29.67	26.84	44.38
Model	Ancient Society & Economy	Keju in Human History	Keju Reform & Abolition	Keju Literature	Keju Definition & Evolution	Keju Materials & Life Sites	/
<i>Closed-Source Models</i>							
Claude-sonnet-4.5-20250929-Thinking	11.09	15.13	12.70	16.27	12.05	11.75	/
GPT-5.2	8.88	3.60	10.74	12.18	10.92	13.00	/
GPT-5.2-Thinking	15.76	13.07	11.07	14.73	14.35	15.00	/
GPT-o3-2025-04-16	13.71	19.40	13.04	20.00	13.45	15.13	/
Gemini-3-Pro-Preview	26.65	26.20	25.30	33.19	23.52	30.38	/
Gemini-3-Pro-Preview-Thinking	30.06	32.47	23.11	31.09	23.81	22.13	/
Qwen3-Max	16.85	10.33	15.63	21.91	17.43	17.50	/
<i>Open-Source Models</i>							
Llama-4-Scout-17B-16E-Instruct	2.09	1.13	1.67	6.55	2.12	4.38	/
gpt-oss-120b	11.88	12.60	9.93	8.18	10.23	14.00	/
gpt-oss-20b	0.91	0.00	0.44	0.00	0.20	0.00	/
Kimi-K2-Thinking	24.68	13.80	17.44	26.36	21.40	28.13	/
GLM-4.6-Thinking	20.59	26.13	24.26	29.64	22.18	26.25	/
Qwen3-14B-Thinking	11.26	9.87	8.52	15.64	10.49	20.50	/
Qwen3-32B-Thinking	13.41	7.13	11.81	16.64	14.42	16.12	/
Qwen3-235B-A22B-Thinking	26.59	23.20	26.48	27.27	27.17	32.75	/
DeepSeek-V3.2	19.47	13.33	12.93	24.91	15.82	20.00	/
DeepSeek-V3.2-Thinking	16.71	13.93	12.37	25.36	14.60	24.13	/
DeepSeek-R1-0528	24.12	23.00	22.11	29.36	25.68	25.75	/

Table 8: Fine-grained performance across specific historical topics. The table reports the evaluation scores of varying LLMs on 13 distinct topics.

- 社，2024年。
- 陈光新编著：《中国筵席宴会大典》，青岛：青岛出版社，1995年。
 - 翟国璋：《中国科举辞典》，南昌：江西教育出版社，2006年。
 - 夏征农：《辞海（中国古代史分册）》，上海：上海辞书出版社，1988年。
 - 杨金鼎主编，上海师范大学古籍整理研究所编：《中国文化史词典》，杭州：浙江古籍出版社，1987年。
 - 吴宗国：《唐代科举制度研究》，北京：北京大学出版社，2022年。
 - 高福顺：《辽朝科举制度》，吉林大学博士学位论文，2008年。
 - 申万里：《元代教育研究》，武汉：武汉大学出版社，2007年。
 - 姚大力：《蒙元制度与政治文化》，北京：北京大学出版社，2011年。
 - 梁庚尧：《北宋的改革与变法》，台北：东方出版中心，2024年。
 - 梁庚尧：《宋代科举社会》，台北：台湾大学出版社，2015年。
 - 祝尚书：《宋代科举与文化》，北京：中华书局，2023年。

Model	Term Interpretation				Fact QA				Historical Reasoning				Average			
	BL	RG	BS	RS	BL	RG	BS	RS	BL	RG	BS	RS	BL	RG	BS	RS
<i>Closed-Source Models</i>																
Claude-Sonnet-4.5-Thinking	1.42	1.76	71.24	12.72	3.36	5.12	71.68	15.21	2.33	6.57	71.42	11.87	2.53	4.76	71.49	12.99
GPT-5.2	2.39	0.32	70.12	11.55	6.01	4.83	72.05	7.68	1.18	4.09	71.78	12.72	3.50	3.46	71.48	11.07
GPT-5.2-Thinking	2.13	1.06	70.37	15.08	7.96	6.24	71.95	11.31	1.81	5.16	71.95	14.95	4.45	4.58	71.55	14.08
GPT-o3-2025-04-16	2.92	1.12	72.15	14.16	14.95	4.68	72.87	15.10	2.77	4.25	72.59	14.79	7.88	3.65	72.60	14.66
Gemini-3-Pro-Preview	0.60	3.24	73.43	29.57	2.61	6.66	74.15	31.86	2.08	7.98	74.13	27.53	1.94	6.27	73.97	29.29
Gemini-3-Pro-Preview-Thinking	0.63	1.65	73.35	24.42	3.70	6.54	74.18	30.47	1.94	6.08	74.03	26.35	2.35	5.16	73.92	26.73
Qwen3-Max	0.00	3.54	74.79	16.35	7.64	7.44	75.14	19.29	4.78	7.96	75.00	17.85	4.77	6.64	75.01	17.71
<i>Open-Source Models</i>																
Llama-4-Scout-17B-16E	0.00	0.43	71.28	2.51	6.21	3.71	73.11	3.46	0.00	4.29	73.18	2.44	2.59	3.09	72.68	2.72
gpt-oss-120b	0.32	0.22	68.69	11.21	2.27	2.31	70.54	7.96	0.73	2.29	70.85	12.07	1.27	1.78	70.18	10.75
gpt-oss-20b	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail
Kimi-K2-Thinking	2.16	3.44	73.25	22.69	6.10	7.21	73.12	22.67	1.61	7.69	73.28	22.93	3.62	6.43	73.20	22.79
GLM-4.6-Thinking	0.48	1.67	71.26	25.93	3.40	5.88	72.69	24.21	1.65	6.73	72.60	23.10	2.09	5.11	72.30	24.32
Qwen3-14B-Thinking	0.34	0.52	71.87	11.08	3.11	4.87	73.29	13.16	1.53	4.34	73.28	11.09	1.89	3.61	72.93	11.61
Qwen3-32B-Thinking	0.43	0.88	71.97	13.49	3.60	5.16	73.25	13.87	1.35	4.81	73.43	14.23	2.06	3.97	72.99	13.89
Qwen3-235B-A22B-Thinking	0.30	3.83	71.70	26.73	1.62	6.23	72.81	30.36	0.99	5.01	72.70	27.95	1.08	5.22	72.50	28.14
DeepSeek-V3.2	0.93	2.64	72.90	16.56	9.25	7.34	73.68	23.30	1.70	7.44	73.80	17.83	4.65	6.20	73.52	18.77
DeepSeek-V3.2-Thinking	0.68	2.57	73.00	16.29	7.99	7.73	73.48	24.47	4.22	6.88	73.62	17.21	4.91	6.16	73.41	18.72
DeepSeek-R1-0528	0.47	2.83	72.22	25.20	3.45	8.71	73.49	31.31	1.14	6.80	73.42	25.55	1.93	6.60	73.15	26.87

Table 9: Detailed results of different LLMs on the T1-T3 tasks. The table presents the performance of representative closed-source and open-source LLMs. The metrics reported are BLEU (BL), ROUGE (RG), BERTScore (BS), and Rubric Score (RS).

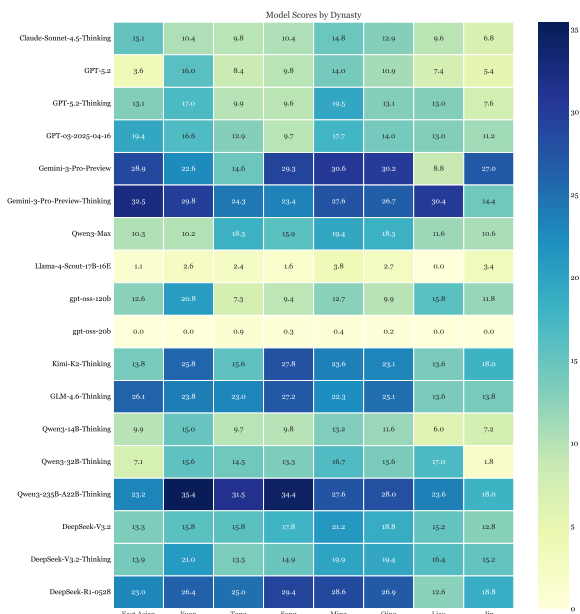


Figure 14: Heatmap of LLM performance across diverse historical periods. Most LLMs perform robustly in mainstream eras (e.g., Tang, Song, Ming) but struggle in low-resource contexts like the Liao and Jin dynasties.

- 陈宝良：《明代儒学生员与地方社会》，北京：中国社会科学出版社，2005年。
- 郭培贵：《明代选举志考论》，北京：中华书局，2006年。
- 龚笃清：《中国八股文史·明代卷》，长沙：岳麓书社，2017年。
- 王颖、黄强：《游戏八股文研究》，武

Model	T1	T2	T3	Average
<i>Closed-Source Models</i>				
Claude-Sonnet-4.5-Thinking	21.40	15.50	21.10	19.33
GPT-5.2	6.40	21.90	16.20	14.83
GPT-5.2-Thinking	13.40	14.30	18.20	15.30
GPT-o3-2025-04-16	21.40	17.20	19.50	19.37
Gemini-3-Pro-Preview	35.60	39.60	27.30	34.17
Gemini-3-Pro-Preview-Thinking	36.90	37.90	24.90	33.23
Qwen3-Max	28.80	28.90	23.10	26.93
<i>Open-Source Models</i>				
Llama-4-Scout-17B-16E-Instruct	4.40	5.00	3.80	4.40
gpt-oss-120b	0.00	7.10	12.90	6.67
gpt-oss-20b	5.00	0.00	0.00	1.67
Kimi-K2-Thinking	29.20	32.00	24.20	28.47
GLM-4.6-Thinking	26.20	36.70	20.30	27.73
Qwen3-14B-Thinking	16.50	14.90	15.40	15.60
Qwen3-32B-Thinking	22.60	20.50	30.20	24.43
Qwen3-235B-A22B-Thinking	31.20	36.90	33.80	33.97
DeepSeek-V3.2	27.80	23.60	16.90	22.77
DeepSeek-V3.2-Thinking	25.20	19.00	17.60	20.60
DeepSeek-R1-0528	33.00	36.60	24.20	31.27
<i>Human</i>				
Historian (Closed-book)	35.00	22.00	31.00	29.33
Historian (Open-book)	74.00	52.00	56.00	60.67

Table 10: Performance comparison between SOTA LLMs and human experts. The evaluation contrasts model performance against *Closed-book* (memory-based) and *Open-book* (research-based) human baselines across three task dimensions (T1-T3).

汉：武汉大学出版社，2015年。

- 安东强：《清代学政规制与皇权体制》，

T1: 对举人大挑进行名词解释

T1: Define *Juren Datiao* (The Grand Selection of Juren).

<p>Golden Description Sample</p> <p>……Rubric2:指出举人大挑是为连续多科落第举子提供入仕机会的考试。[Point out that <i>Juren Datiao</i> (Grand Selection of <i>Juren</i>) was a selection process designed to provide opportunities for entry into officialdom for <i>Juren</i> who had failed the Metropolitan Examination for several consecutive sessions.]</p> <p>……Rubric4:明确举人大挑的考试内容为仪表外貌与应对。[Specify that the assessment criteria for <i>Juren Datiao</i> focused on the candidates' physical appearance and their manner of verbal responsiveness (conversational poise).]</p> <p>……</p>	<p>Gemini-3-Pro-Preview *Score:74</p>	<p>……凡是连续三次（一说四次，后定为三科）参加会试而未能考中进士的举人，均有资格报名参加“大挑”。……考官主要根据考生的相貌、体格、应对进退来定去留……[All <i>Juren</i> who had participated in the Metropolitan Examination for three consecutive sessions (some sources say four, but it was later fixed at three) without attaining the <i>Jinshi</i> degree were eligible to apply for the <i>Datiao</i>. The examiners determined the selection or rejection of candidates primarily based on their physical appearance, physique, and manner of conduct and verbal responsiveness.]</p>	✓
	<p>Qwen3-235B-A22B-Thinking *Score:78</p>	<p>举人大挑……对象是会试多次落第（通常要求三科以上）且未被选用的举人。……选拔的核心标准是相貌、身材、气质和应对谈吐……[<i>Juren Datiao</i>……was targeted at <i>Juren</i> who had failed the Metropolitan Examination multiple times (typically required to have failed at least three sessions) and had not yet been appointed to any official posts. The……core criteria for selection were physical appearance, stature, temperament, and verbal responsiveness (the manner of conversation and poise)……]</p>	✓
	<p>DeepSeek-R1-0528 *Score:87</p>	<p>……并非所有举人都能参加，通常需要满足参加会试达到一定次数（如三科不中）或达到一定年龄。……选拔核心是面试，考察“形貌”和“应对”……[Participation was not open to all <i>Juren</i>, it was restricted to those who had failed the Metropolitan Examination a certain number of times (typically three sessions) or had reached a specific age.……The core criteria for evaluation were “physical appearance” and “verbal responsiveness” (poise and eloquence in answering).……]</p>	✓
	<p>GPT-5.2-Thinking *Score:26</p>	<p>清代……定期在京对已中举（举人）但未成进士、长期候补者进行集中“挑选”（考察才具、文理等并分等第）……的一种特别选拔方式……[This was a special selection mechanism conducted periodically in the capital for <i>Juren</i> who had not yet attained the <i>Jinshi</i> (Metropolitan Graduate) degree and had remained on the waiting list for official appointments for an extended period. The process involved a centralized evaluation of the candidates' aptitude, literary proficiency, and logic, followed by their classification into different grades for appointment.……]</p>	✗

T1: 对清代的科举中的磨勘进行名词解释。

T1: Define the *Mokan* (Post-examination Audit) in the Qing Dynasty imperial examinations.

<p>Golden Description Sample</p> <p>……Rubric2:回答清代磨勘的流程，即乡、会试揭榜后将中式者的试卷送往朝廷，朝廷派官对试卷进行检查。[State the process of <i>Mokan</i> (post-examination audit) in the Qing Dynasty: after the results of the Provincial (<i>Xiangshi</i>) and Metropolitan (<i>Huishhi</i>) examinations were announced, the answer scripts of the successful candidates were sent to the imperial court, where the court would appoint officials to conduct a thorough inspection of the papers.]</p> <p>……</p>	<p>Gemini-3-Pro-Preview *Score:50</p>	<p>……时间通常在乡试或会试放榜之后进行。由朝廷指派“磨勘官”（通常由翰林院官员、御史等担任）。对象主要针对中式（考中）者的试卷。……[The process usually takes place after the results of the Provincial Examination (<i>Xiangshi</i>) or Metropolitan Examination (<i>Huishhi</i>) are announced. The Imperial Court appoints “Audit Officials” (<i>Mokan Guan</i>), who are typically drawn from the ranks of the Hanlin Academy and the Censorate (Imperial Censors). The audit is primarily directed at the examination scripts of the successful candidates (those who have passed).……]</p>	✓
	<p>Qwen3-235B-A22B-Thinking *Score:50</p>	<p>……通常在乡试、会试放榜之后进行（殿试后也有类似程序但相对简化）。中举/中式的名单虽已公布，但最终功名需待磨勘通过后方能正式确认。……[This process was typically conducted after the results of the Provincial (<i>Xiangshi</i>) and Metropolitan (<i>Huishhi</i>) examinations were announced (a similar, albeit simplified, procedure also followed the Palace Examination). Although the names of the successful candidates had already been publicized, their final academic degrees and honors could only be officially confirmed once they had successfully passed the <i>Mokan</i> audit.……]</p>	✓
	<p>DeepSeek-R1-0528 *Score:50</p>	<p>“磨勘”是在乡试、会试等重大考试放榜后，由中央专门机构（主要是礼部）对中式举人、贡士的试卷进行复查，以及对相关考官工作进行审核的制度。……[<i>Mokan</i> (Post-examination Audit) specifically referred to the system in which, after the results of the Provincial Examination (<i>Xiangshi</i>) and Metropolitan Examination (<i>Huishhi</i>) were announced, specialized central agencies (primarily the Ministry of Rites) would conduct a thorough re-evaluation of the examination scripts of successful candidates (<i>Juren</i> and <i>Gongshi</i>) and audit the performance of the relevant examination officials.……]</p>	✓
	<p>GPT-5.2-Thinking *Score:33</p>	<p>……在主考官初阅卷、拟定名次之后，由另派官员（称“磨勘官”）对已取中的试卷及录取名册进行逐卷核查。……[After the Chief Examiners have completed the initial grading, finalized the selection of scripts, and drafted the rankings, separately appointed officials known as “Audit Officials” (<i>Mokan Guan</i>) perform a script-by-script verification of the successful candidates' examination papers and the official recruitment registry.……]</p>	✗

Figure 15: Case study on Task T1 (Part I): Qualitative Results.

T1: 对知贡举进行名词解释。

T1: Provide a definition for *Zhigongju* (Chief Examiner)

<p>Golden Description Sample</p> <p>……Rubric5:提到知贡举的历史沿革: 说明知贡举最早出现的时间在唐朝。[In discussing the historical evolution of the <i>Zhigongju</i> (Chief Examiner), it should be noted that the position first emerged during the Tang Dynasty.]</p> <p>……</p>	<p>Gemini-3-Pro-Preview</p> <p>*Score:67</p>	<p>……唐代科举常科最初由吏部考功员外郎主持, 唐玄宗开元年间改由礼部侍郎主持。此时的主考官通常被称为“知贡举”。……[……The regular examinations (<i>Changke</i>) of the Tang Dynasty were presided over by the <i>Kaogong Yuanwailang</i> of the Ministry of Personnel. During the Kaiyuan era of Emperor Xuanzong, the responsibility was shifted to the <i>Libu Shilang</i> (Vice Minister of Rites). At this time, the chief examiner was commonly referred to as the “Zhigongju”.……]</p>	✓
	<p>Qwen3-235B-A22B-Thinking</p> <p>*Score:67</p>	<p>……起源与发展: (知贡举) 此职始设于唐朝。早期常由吏部考功员外郎主持, 后因地位过低、易受权贵干扰, 改由礼部侍郎 (后固定为礼部尚书或侍郎) 担任知贡举。……[……Origins and Development: This position was first established during the Tang Dynasty. Initially, the examinations were administered by the Vice Director of the Bureau of Merit (<i>Kaogong Yuanwailang</i>) under the Ministry of Personnel.……]</p>	✓
	<p>DeepSeek-R1-0528</p> <p>*Score:50</p>	<p>……唐代: 是“知贡举”制度形成和普遍使用的时期。通常由礼部侍郎担任。……[……The Tang Dynasty was the period during which the <i>Zhigongju</i> system was officially formed and came into widespread use. Typically held by the Vice Minister of Rites.……]</p>	✓
	<p>GPT-5.2-Thinking</p> <p>*Score:33</p>	<p>……指由皇帝临时任命、主持“贡举”的官员……大体相当于后世所称的主考官 (多见于唐宋, 后代沿用或有相近称谓)。……[……<i>Zhigongju</i>. Refers to officials temporarily appointed by the Emperor to preside over the civil service examinations (<i>Gongju</i>). This role was roughly equivalent to what was called the “Chief Examiner” in later generations (primarily seen in the Tang and Song dynasties, with the title or similar designations continuing in subsequent eras).……]</p>	✗

T1: 对会试进行名词解释。

T1: Provide a definition for *Huishi* (Metropolitan Examination)

<p>Golden Description Sample</p> <p>……Rubric4:答出会试在金朝、元朝、明清的变化, 主要包括中额数的变化、考试时间的变化、会试在科举考试中的级别变化等。[Describe the evolution of the Metropolitan Examination (<i>Huishi</i>) during the Jin, Yuan, Ming, and Qing dynasties, primarily focusing on changes in admission quotas, examination timing, and its hierarchical status within the imperial examination system.]</p> <p>……</p>	<p>Gemini-3-Pro-Preview</p> <p>*Score:36</p>	<p>基本定义: 会试是科举考试中, 继“乡试”(省级考试) 之后的第二级考试……是由“举人”晋升为“贡士”的全国性选拔考试。(全文未提及 rubric1 内容) [Basic Definition: The Metropolitan Examination (<i>Huishi</i>) is the second-tier examination in the imperial civil service system, following the “Provincial Examination” (<i>Xiangshi</i>). It is a nationwide selection process through which <i>Juren</i> (Provincial Graduates) advance to become <i>Gongshi</i> (Tribute Scholars). (The entire text makes no mention of the content related to rubric1.)]</p>	✗
	<p>Qwen3-235B-A22B-Thinking</p> <p>*Score:36</p>	<p>……“唐宋称‘会试’”……“会试”一词定型于明清; ……“唐宋类似考试称‘省试’ (由尚书省主持)”。……[……Regarding the claim that “the Tang and Song called it <i>Huishi</i>”……the term <i>Huishi</i> (Metropolitan Examination) was actually standardized during the Ming and Qing dynasties.……In the Tang and Song, the equivalent examination was known as <i>Shengshi</i> (Departmental Examination), as it was administered by the Department of State Affairs (<i>Shangshu Sheng</i>).……]</p>	✗
	<p>DeepSeek-R1-0528</p> <p>*Score:36</p>	<p>……会试是明清时期 (其制度成熟于明清, 但可追溯至唐宋) 在京城由礼部主持的全国性科举考试, 于乡试次年的春天举行。……[……The Metropolitan Examination (<i>Huishi</i>) was a nationwide imperial examination held in the capital during the Ming and Qing dynasties (the system reached maturity during this era, though its origins can be traced back to the Tang and Song).……]</p>	✗
	<p>GPT-5.2-Thinking</p> <p>*Score:36</p>	<p>会试: 中国古代科举制度中的一级考试, 通常在京城由礼部主持 (又称“礼部试”“春闹”), ……(全文未提及 rubric1 内容) [Metropolitan Examination (<i>Huishi</i>)]</p> <p>A major level of the imperial examination system in ancient China, the <i>Huishi</i> was typically held in the capital and administered by the Ministry of Rites (hence also known as the “Ministry Examination” or “Spring Examination”). (The entire text makes no mention of the content related to rubric1.)]</p>	✗

Figure 16: Case study on Task T1 (Part II): Qualitative Results.

T2: 明代如何惩治科场违规、作弊的考生?

T2: How did the Ming Dynasty punish candidates for examination irregularities and cheating?

Model	Score	Qualitative Result
Gemini-3-Pro-Preview	*Score:63	✓
Qwen3-235B-A22B-Thinking	*Score:14	✗
DeepSeek-R1-0528	*Score:33	✗
GPT-5.2-Thinking	*Score:0	✗

Golden Description Sample

……Rubric2:答出明代对科场作弊的学生有以下惩戒方式:首先是当场“扶出”。[The Ming Dynasty had the following punishment methods for students who cheated in the imperial examination: first, they were “*fuchu*” on the spot.]
 Rubric3:答出“扶出”即取消各种违规考生参加该科考试的资格。[Answering “*fuchu*” means revoking the eligibility of all non compliant candidates to take the exam in that subject.]

T2: 简述宋代的殿试制度

T2: Briefly describe the palace examination system in the Song Dynasty

Model	Score	Qualitative Result
Gemini-3-Pro-Preview	*Score:18	✓
Qwen3-235B-A22B-Thinking	*Score:25	✓
DeepSeek-R1-0528	*Score:39	✓
GPT-5.2-Thinking	*Score:14	✗

Golden Description Sample

……Rubric25:答出来登科题名分为及第、出身、同出身。[In the Song Dynasty, the titles conferred upon candidates who passed the imperial examinations (*dengke*) were categorized into three grades: Jidi (Passed with Highest Distinction), Chushen (Possessing the Status), and Tong-chushen (Equivalent to the Status).]

Figure 17: Case study on Task T2 (Part I): Qualitative Results.

T2: 回答唐代常举科举自地方至中央的完整运行机制。

T2: Explain the complete operational mechanism of the *Changju* (Regular Imperial Examinations) in the Tang Dynasty, from the local level to the central government.

Golden Description Sample													
<p>……Rubric10: 答出明经、进士通过各项审核后, 于来年春天参加考试, 故亦称春闱, 时间一般在正月。 [State that candidates of the Mingjing and Jinshi categories, after passing all necessary audits and reviews, would sit for the examination in the spring of the following year. Consequently, it was also known as the "Spring Examination" (Chunwei), with the timing generally set in the first lunar month.]</p> <p>……</p>	<table border="1"> <tr> <td> <p>Gemini-3-Pro-Preview *Score:11</p> </td> <td> <p>……通常在次年春季 (二左右), 故称“春闱”。……[……It was typically held in the spring of the following year (around the second lunar month), and was thus known as the "Spring Examination" (Chunwei). ……]</p> </td> <td>X</td> </tr> <tr> <td> <p>Qwen3-235B-A22B-Thinking *Score:41</p> </td> <td> <p>……中央环节: 尚书省省试 (“礼部试”)。次年正月 (解试次年春) 赴京与报到。……[……The Central Stage: The Departmental Examination (Shengshi) of the Department of State Affairs (also known as the "Ministry of Rites Examination"). Candidates would travel to the capital to report for registration in the first lunar month of the following year (the spring following the local qualifying examination (Jieshi). ……]</p> </td> <td>✓</td> </tr> <tr> <td> <p>DeepSeek-R1-0528 *Score:35</p> </td> <td> <p>……中央考试: 省试。时间: 每年正月 (后固定于二至三月)。……[……Central Examination: The Departmental Examination (Shengshi). Timing: Held annually in the first lunar month (later fixed between the second and third lunar months). ……]</p> </td> <td>✓</td> </tr> <tr> <td> <p>GPT-5.2-Thinking *Score:21</p> </td> <td> <p>……中央核心环节: 礼部贡举 (“礼部试/省试”), 礼部官主持, 设副主考与同考官; 专科可兼取相关部门官员协助。…… (未涉及 rubric 要求的时间说明。) [……The core central stage: The Ministry of Rites Examinations (Shengshi). These were presided over by officials from the Ministry of Rites, with Deputy Examiners and Co-examiners appointed. …… (Missing required time-related information in rubric.)]</p> </td> <td>X</td> </tr> </table>	<p>Gemini-3-Pro-Preview *Score:11</p>	<p>……通常在次年春季 (二左右), 故称“春闱”。……[……It was typically held in the spring of the following year (around the second lunar month), and was thus known as the "Spring Examination" (Chunwei). ……]</p>	X	<p>Qwen3-235B-A22B-Thinking *Score:41</p>	<p>……中央环节: 尚书省省试 (“礼部试”)。次年正月 (解试次年春) 赴京与报到。……[……The Central Stage: The Departmental Examination (Shengshi) of the Department of State Affairs (also known as the "Ministry of Rites Examination"). Candidates would travel to the capital to report for registration in the first lunar month of the following year (the spring following the local qualifying examination (Jieshi). ……]</p>	✓	<p>DeepSeek-R1-0528 *Score:35</p>	<p>……中央考试: 省试。时间: 每年正月 (后固定于二至三月)。……[……Central Examination: The Departmental Examination (Shengshi). Timing: Held annually in the first lunar month (later fixed between the second and third lunar months). ……]</p>	✓	<p>GPT-5.2-Thinking *Score:21</p>	<p>……中央核心环节: 礼部贡举 (“礼部试/省试”), 礼部官主持, 设副主考与同考官; 专科可兼取相关部门官员协助。…… (未涉及 rubric 要求的时间说明。) [……The core central stage: The Ministry of Rites Examinations (Shengshi). These were presided over by officials from the Ministry of Rites, with Deputy Examiners and Co-examiners appointed. …… (Missing required time-related information in rubric.)]</p>	X
<p>Gemini-3-Pro-Preview *Score:11</p>	<p>……通常在次年春季 (二左右), 故称“春闱”。……[……It was typically held in the spring of the following year (around the second lunar month), and was thus known as the "Spring Examination" (Chunwei). ……]</p>	X											
<p>Qwen3-235B-A22B-Thinking *Score:41</p>	<p>……中央环节: 尚书省省试 (“礼部试”)。次年正月 (解试次年春) 赴京与报到。……[……The Central Stage: The Departmental Examination (Shengshi) of the Department of State Affairs (also known as the "Ministry of Rites Examination"). Candidates would travel to the capital to report for registration in the first lunar month of the following year (the spring following the local qualifying examination (Jieshi). ……]</p>	✓											
<p>DeepSeek-R1-0528 *Score:35</p>	<p>……中央考试: 省试。时间: 每年正月 (后固定于二至三月)。……[……Central Examination: The Departmental Examination (Shengshi). Timing: Held annually in the first lunar month (later fixed between the second and third lunar months). ……]</p>	✓											
<p>GPT-5.2-Thinking *Score:21</p>	<p>……中央核心环节: 礼部贡举 (“礼部试/省试”), 礼部官主持, 设副主考与同考官; 专科可兼取相关部门官员协助。…… (未涉及 rubric 要求的时间说明。) [……The core central stage: The Ministry of Rites Examinations (Shengshi). These were presided over by officials from the Ministry of Rites, with Deputy Examiners and Co-examiners appointed. …… (Missing required time-related information in rubric.)]</p>	X											

T3: 论述宋代的特奏名制度。

T3: Discuss the Tezouming (Special List) System of the Song Dynasty.

Golden Description Sample													
<p>……Rubric4: 答出开宝三年 (970) 设立特奏名制度, 仁宗朝特奏名成为定制。 [State that the Tezouming (Special List) system was established in the third year of the Kaibao era (970), and became a formalized institution during the reign of Emperor Renzong.]</p> <p>……</p>	<table border="1"> <tr> <td> <p>Gemini-3-Pro-Preview *Score:70</p> </td> <td> <p>……特奏名制度始于宋初, 太祖、太宗朝并无定制, 多由皇帝临时起意。……到了宋仁宗时期, 规定更加细化, “试礼部十上”或“年五十以上”者可由礼部具名特奏。……[The "Tezouming" system originated in the early Song Dynasty. During the reigns of Emperors Taizu and Taizong, conferments were largely based at the emperor's discretion. By the reign of Emperor Renzong, the regulations became more granular. Candidates who had "sat for the Ministry of Rites examination ten times" or were "over fifty years of age" could be formally listed by the Ministry of Rites.]</p> </td> <td>✓</td> </tr> <tr> <td> <p>Qwen3-235B-A22B-Thinking *Score:81</p> </td> <td> <p>……初创与谨慎: 首开于太祖建隆三年 (962年), 人数极少, 体现恩典性质。……制度化与扩大: 仁宗景祐元年 (1034年) 诏书标志制度成熟。人数显著增加 (如仁宗朝平均每科特奏名约300人, 接近正奏名)。……[……It was first introduced in the third year of the Jianlong era of Emperor Taizu (962 AD). ……The imperial edict in the first year of the Jingyou era of Emperor Renzong (1034 AD) marked the maturity of the system. ……]</p> </td> <td>X</td> </tr> <tr> <td> <p>DeepSeek-R1-0528 *Score:81</p> </td> <td> <p>……北宋真宗景德二年 (1005年) 正式确立基本框架: 进士科五举 (参加五次省试)。……[……The basic framework was formally established in the second year of the Jingde era (1005 AD) under Emperor Zhenzong of the Northern Song: [it required] five attempts (Wujiu) for the Jinshi category, meaning participation in five Departmental Examinations (Shengshi). ……]</p> </td> <td>X</td> </tr> <tr> <td> <p>GPT-5.2-Thinking *Score:14</p> </td> <td> <p>……制度从偶行到定制, 再到反复收紧。北宋前期多带临时恩赐性质, 尚未完全稳定成例; ……北宋中后期逐步“常例化”。……[……The system evolved from an ad hoc practice to a fixed institution, followed by periods of repeated tightening. In the early Northern Song, it was largely characterized by temporary imperial grace and had not yet fully stabilized into an established precedent; ……by the mid-to-late Northern Song, it gradually became "routinized". ……]</p> </td> <td>X</td> </tr> </table>	<p>Gemini-3-Pro-Preview *Score:70</p>	<p>……特奏名制度始于宋初, 太祖、太宗朝并无定制, 多由皇帝临时起意。……到了宋仁宗时期, 规定更加细化, “试礼部十上”或“年五十以上”者可由礼部具名特奏。……[The "Tezouming" system originated in the early Song Dynasty. During the reigns of Emperors Taizu and Taizong, conferments were largely based at the emperor's discretion. By the reign of Emperor Renzong, the regulations became more granular. Candidates who had "sat for the Ministry of Rites examination ten times" or were "over fifty years of age" could be formally listed by the Ministry of Rites.]</p>	✓	<p>Qwen3-235B-A22B-Thinking *Score:81</p>	<p>……初创与谨慎: 首开于太祖建隆三年 (962年), 人数极少, 体现恩典性质。……制度化与扩大: 仁宗景祐元年 (1034年) 诏书标志制度成熟。人数显著增加 (如仁宗朝平均每科特奏名约300人, 接近正奏名)。……[……It was first introduced in the third year of the Jianlong era of Emperor Taizu (962 AD). ……The imperial edict in the first year of the Jingyou era of Emperor Renzong (1034 AD) marked the maturity of the system. ……]</p>	X	<p>DeepSeek-R1-0528 *Score:81</p>	<p>……北宋真宗景德二年 (1005年) 正式确立基本框架: 进士科五举 (参加五次省试)。……[……The basic framework was formally established in the second year of the Jingde era (1005 AD) under Emperor Zhenzong of the Northern Song: [it required] five attempts (Wujiu) for the Jinshi category, meaning participation in five Departmental Examinations (Shengshi). ……]</p>	X	<p>GPT-5.2-Thinking *Score:14</p>	<p>……制度从偶行到定制, 再到反复收紧。北宋前期多带临时恩赐性质, 尚未完全稳定成例; ……北宋中后期逐步“常例化”。……[……The system evolved from an ad hoc practice to a fixed institution, followed by periods of repeated tightening. In the early Northern Song, it was largely characterized by temporary imperial grace and had not yet fully stabilized into an established precedent; ……by the mid-to-late Northern Song, it gradually became "routinized". ……]</p>	X
<p>Gemini-3-Pro-Preview *Score:70</p>	<p>……特奏名制度始于宋初, 太祖、太宗朝并无定制, 多由皇帝临时起意。……到了宋仁宗时期, 规定更加细化, “试礼部十上”或“年五十以上”者可由礼部具名特奏。……[The "Tezouming" system originated in the early Song Dynasty. During the reigns of Emperors Taizu and Taizong, conferments were largely based at the emperor's discretion. By the reign of Emperor Renzong, the regulations became more granular. Candidates who had "sat for the Ministry of Rites examination ten times" or were "over fifty years of age" could be formally listed by the Ministry of Rites.]</p>	✓											
<p>Qwen3-235B-A22B-Thinking *Score:81</p>	<p>……初创与谨慎: 首开于太祖建隆三年 (962年), 人数极少, 体现恩典性质。……制度化与扩大: 仁宗景祐元年 (1034年) 诏书标志制度成熟。人数显著增加 (如仁宗朝平均每科特奏名约300人, 接近正奏名)。……[……It was first introduced in the third year of the Jianlong era of Emperor Taizu (962 AD). ……The imperial edict in the first year of the Jingyou era of Emperor Renzong (1034 AD) marked the maturity of the system. ……]</p>	X											
<p>DeepSeek-R1-0528 *Score:81</p>	<p>……北宋真宗景德二年 (1005年) 正式确立基本框架: 进士科五举 (参加五次省试)。……[……The basic framework was formally established in the second year of the Jingde era (1005 AD) under Emperor Zhenzong of the Northern Song: [it required] five attempts (Wujiu) for the Jinshi category, meaning participation in five Departmental Examinations (Shengshi). ……]</p>	X											
<p>GPT-5.2-Thinking *Score:14</p>	<p>……制度从偶行到定制, 再到反复收紧。北宋前期多带临时恩赐性质, 尚未完全稳定成例; ……北宋中后期逐步“常例化”。……[……The system evolved from an ad hoc practice to a fixed institution, followed by periods of repeated tightening. In the early Northern Song, it was largely characterized by temporary imperial grace and had not yet fully stabilized into an established precedent; ……by the mid-to-late Northern Song, it gradually became "routinized". ……]</p>	X											

Figure 18: Case study on Task T2 (Part II) and T3 (Part I): Qualitative Results.

T3: 元代科举有哪些特点?

T3: What were the characteristics of the imperial examination system in the Yuan Dynasty?

<p>Golden Description Sample</p> <p>…… Rubric7: 答出元代科举左榜是针对汉人、南人。 [State that the Left List (<i>Zuobang</i>) of the Yuan Dynasty imperial examinations was specifically designated for the Han people (<i>Hanren</i>) and Southerners (<i>Nanren</i>).]</p>	<p>Gemini-3-Pro-Preview *Score:41</p>	<p>……左右榜制度：考试分为“右榜”（供蒙古人、色目人）和“左榜”（供汉人、南人）。元代尚右，右榜地位高于左榜。……[……The examination was divided into the “Right List” (known as <i>Youbang</i>, designated for Mongols and <i>Semu</i> people) and the “Left List” (known as <i>Zuobang</i>, designated for Han people and Southerners)……]</p>	✓
	<p>Qwen3-235B-A22B-Thinking *Score:8</p>	<p>……分榜录取，名额不均。左榜（蒙古、色目人）：试题较易，录取名额占总数50%。右榜（汉人、南人）……[……The Left List (for Mongols and <i>Semu</i>): The examination questions were easier, and the recruitment quota accounted for 50% of the total. The Right List (for Han and Southerners): The questions were more difficult……]</p>	✗
	<p>DeepSeek-R1-0528 *Score:45</p>	<p>……考试分左右两榜。左榜供汉人（原金朝统治区及较早归附的汉族、契丹、女真等）、南人（原南宋统治区的汉族）应试。……[……The examination was divided into Left and Right lists. The Left List was for Han people (referring to those from the former Jin Dynasty territory and early-submitting groups like the Han, Khitan, and Jurchen) and Southerners (referring to the Han people from the former Southern Song territory) to sit for the exam.……]</p>	✓
	<p>GPT-5.2-Thinking *Score:28</p>	<p>……元代科举最具辨识度的制度特征：以“蒙古、色目、汉人、南人”四类身份为基础，分卷命题或至少分定录取名额。……未提及rubric要求的榜内容。[……The most recognizable institutional characteristic of the Yuan Dynasty examinations: Based on the four ethnic identities—“Mongol, <i>Semu</i>, Han, and Southerner”—the system featured differentiated exam questions or, at minimum, separately fixed recruitment quotas.……(No mention of the “list” content required by the rubric.)]</p>	✗

T4: 假设你是一位清朝道光甲辰科（道光二十四年）的科举会试考生，请你严格按照八股文的写作方式写一篇科场八股文，如果遇到需要避讳的字，以拼音标注。字数不要超过700字。请直接作答，仅输出你写的八股文，不要输出其他内容。题目是：“下学而上达，知我者其天乎。” [Assume you are a candidate for the *Jiachén* year Metropolitan Examination during the Daoguang era of the Qing Dynasty (the 24th year of Daoguang, 1844). Please write an Eight-Legged Essay (*Baguwen*) strictly according to its formal requirements. If you encounter characters that are subject to imperial naming taboos, annotate them using *pinyin*. The length should not exceed 700 characters. Please answer directly, providing only the essay itself without any other content. The topic is: “下学而上达，知我者其天乎。”]

<p>Golden Description Sample</p> <p>…… Rubric: ……减分项3: 未对“玄”“焯”“胤”“禛”“弘”“历”“顯”“琰”“璉”“旻”“宁”“丘”以拼音标注避讳。（在道光二十四年科举考试中，不允许直接使用以上避讳字。）</p>	<p>Gemini-3-Pro-Preview *Score:84</p>	<p>……及其深造乎精微，默契乎hóng化，一以贯之，同符造物，确信希天之旨，更有非言语文字之所能该者。……圣人浑然与天为一，又何必求知于人，而ning耐以待后世子云之知我哉？</p>	✓
	<p>Qwen3-235B-A22B-Thinking *Score:0</p>	<p>……夫下学上达，孔门一贯之宗。知我者其天乎，盖言至道玄微，惟苍昊默鉴，非谗闻小识所可拟议。……</p>	✗
	<p>DeepSeek-R1-0528 *Score:84</p>	<p>……自洒扫应对以达礼乐刑hóng（弘）政，自子臣弟友以通天地民物，步步皆实地，而即步步有天机。……韦编之绝，hóng（弘）道弥坚，此际之力即天力也。……</p>	✓
<p>GPT-5.2-Thinking *Score:94</p>	<p>夫子言学必“下”而后“上”，言“知”惟天乃能。下学者，谨乎日用人伦；上达者，通乎性命天理。至诚既具，不求人知，故曰：“知我者其天乎！”盖学不自卑近，则高明无由而至；道不本躬行，则天命终隔而远。圣贤所以阶梯其学者，先立其近以积其远，后穷其理以会其天也。试观夫子在洙泗，删述六经，诲人不倦。弟子或望其显用，而夫子独叹：“知我者其天乎！”非厌世也，惟见学至上达，非俗耳所及耳。故其下学也，事亲则尽孝，事长则尽弟，处众则守信，莅事则守敬。由一言一行而检身，由一日一月而积功，务在切近，不敢躐等。及其上达也，由礼以通乐，由学以入道，推人伦而见性，穷事理而契天。内省无欺，则天理昭然；外行有恒，则德业日进。于是毁誉不足移其心，穷达不足易其守。人知我者，或知其言之善；天知我者，兼知其心之诚。心诚则无隐于上，行直则自与天合。今之士子，读书为文，苟务高谈而忽日用，是失下学之基；苟徇声名而求知人，是背至上达之旨。愿循夫子之训，反躬实行，以渐至于通天之境，则不求闻达而天自鉴矣。</p>	✓	

Figure 19: Case study on Task T3 (Part II) and T4 (Part I): Qualitative Results.

T4: 假设你是一位清朝康熙二十一年壬戌科的科举会试考生，请你严格按照八股文的写作方式写一篇科场八股文，如果遇到需要避讳的字，以拼音标注。字数不要超过650字。请直接作答，仅输出你写的八股文，不要输出其他内容。*圣人治天下，使有菽粟如水火。菽粟如水火，而民焉有不仁者乎？*

[Assume you are a candidate for the *Renxu* Metropolitan Examination during the 21st year of the Kangxi era (1682) of the Qing Dynasty. Please write an Eight-Legged Essay (*Baguwen*) strictly according to its formal requirements. If you encounter characters that are subject to imperial naming taboos, annotate them using *pinyin*. The length should not exceed 650 characters. Please answer directly, providing only the essay itself without any other content. The topic is: "圣人治天下，使有菽粟如水火。菽粟如水火，而民焉有不仁者乎？"]

Golden Description Sample

…… **Rubric:** ……加分项11: 出题在束股之后，大结之前。一般为散句单行。目的是为了跳出四比八股，作更深入的联想和议论，为下一步结束全文作过渡。

T4: 假设你是一位清朝嘉庆丁丑科（嘉庆二十二年）的科举考生，请你严格按照八股文的写作方式写一篇科场八股文，如果遇到需要避讳的字，以拼音标注。字数不要超过700字。请直接作答，仅输出你写的八股文，不要输出其他内容。题目是：子曰：为政以德。

[Assume you are an imperial examination candidate during the *Dingchou* year of the Jiaqing era in the Qing Dynasty (the 22nd year of Jiaqing, 1817). Please write an Eight-Legged Essay (*Baguwen*) in strict accordance with the formal requirements of the examination hall. If you encounter characters that are subject to imperial naming taboos, annotate them with *pinyin*. The word count should not exceed 700 characters. Please answer directly and output only the essay itself; do not output any other content. The topic is: "子曰：为政以德。"]

Golden Description Sample

…… **Rubric:** ……减分项3: 未对“玄”“焯”“胤”“禛”“弘”“历”“颙”“琰”“璉”“丘”以拼音标注避讳。（在嘉庆二十二年科举考试中，不允许直接使用以上避讳字。）

<p>Gemini-3-Pro-Preview *Score:100</p>	<p>(束股) ……是则富而后教，理固宜然，而给足其求，即所以挽回其气。 …… (出题)</p>	<p>✓</p>
<p>Qwen3-235B-A22B-Thinking *Score:94</p>	<p>……束股：是故足食为仁政之基，养民即所以教民，其理昭然。大结：伏惟我皇上xuan ye, 承天景命，勤恤民隐。愿推此道，使天下永享丰年，则仁风遍于寰宇，而圣治光于万世矣。民焉有不仁者乎？其效可翘足而待也。 (缺少出题)</p>	<p>✗</p>
<p>DeepSeek-R1-0528 *Score:76</p>	<p>……束股：是知化民之道，莫先于富民；富民之后，乃可教仁。若舍菽粟而空谈教令，则民虽欲仁，而势有所不能。愿在上者体圣人之心，使菽粟如水火，则不仁之问，可不待而自息矣。(直接以束股结束，没有出题)</p>	<p>✗</p>
<p>GPT-5.2-Thinking *Score:94</p>	<p>……(束比) 故知养民即所以教民，足食实为兴仁之本。圣人不侈言仁义，而仁义自孚于粒食之中。盖元气充则百体舒，仓廩实则礼义生。此尧舜所以命覆为急，而《豳风》所以重农事也。(大结) 我皇上矜念民依，宵旰图治。蠲租赐复，德泽如春。比岁以来，嘉禾叠瑞，红朽相望。将见户习敦庞，俗登熙皞。仁风翔洽，其在斯时乎！……(缺少出题)</p>	<p>✗</p>
<p>Gemini-3-Pro-Preview *Score:88</p>	<p>……是一念之恪守，即万化之枢机。毋务纷更，而xuan德自能默运；是一身之仪型，即四海之观瞻。不事作为，而懿德自不仪形。……为政者其务修诸己哉。</p>	<p>✓</p>
<p>Qwen3-235B-A22B-Thinking *Score:0</p>	<p>……故曰：得乎丘民而为天子，失德者失其位。德者本也，财者末也，本乱而未治者否矣。……</p>	<p>✗</p>
<p>DeepSeek-R1-0528 *Score:94</p>	<p>圣论王猷，首揭德以为政本。盖德者，天心之流行，而政之所由立也。夫子揭其要，万世之纲维在兹矣。……圣人以“为政”垂训，其示万世以心法也夫！钦哉！</p>	<p>✓</p>
<p>GPT-5.2-Thinking *Score:100</p>	<p>圣门示治，一言曰：“为政以德。”万世行政，悉从此出。政所以正人，德所以正己。己正则人正，身修则化行。……在上者，日省其私以培其德；居官者，时戒其利以全其节。推而广之，则家齐、国治、天下平。</p>	<p>✓</p>

Figure 20: Case study on Task T4 (Part II): Qualitative Results.

Model	Open-Source	# Params	Institution	Deployment	Domain
Claude-Sonnet-4.5-Thinking (Anthropic, 2025)	No	-	Anthropic	Official API	General
GPT-5.2 (OpenAI, 2025a)	No	-	OpenAI	Official API	General
GPT-5.2-Thinking (OpenAI, 2025a)	No	-	OpenAI	Official API	General
GPT-o3 (OpenAI, 2025b)	No	-	OpenAI	Official API	General
Gemini-3-Pro-Preview (Google DeepMind, 2025)	No	-	Google DeepMind	Official API	General
Gemini-3-Pro-Preview-Thinking (Google DeepMind, 2025)	No	-	Google DeepMind	Official API	General
Qwen3-Max (Qwen Team, 2025)	No	> 1T	Alibaba	Official API	General
Llama-4-Scout-17B-16E (Meta Llama Team, 2025)	Yes	109B (17B active)	Meta	Locally Load	General
gpt-oss-120b (OpenAI et al., 2025)	Yes	116.8B (5.1B active)	OpenAI	Locally Load	General
gpt-oss-20b (OpenAI et al., 2025)	Yes	20.9B (3.6B active)	OpenAI	Locally Load	General
Kimi-K2-Thinking (Team et al., 2025b)	Yes	1T (32B active)	Moonshot AI	Locally Load	General
GLM-4.6-Thinking (Team et al., 2025a)	Yes	335B (32B active)	Zhipu AI	Locally Load	General
Qwen3-14B-Thinking (Yang et al., 2025)	Yes	14B	Alibaba	Locally Load	General
Qwen3-32B-Thinking (Yang et al., 2025)	Yes	32B	Alibaba	Locally Load	General
Qwen3-235B-A22B-Thinking (Yang et al., 2025)	Yes	235B (22B active)	Alibaba	Locally Load	General
DeepSeek-V3.2 (DeepSeek-AI et al., 2025)	Yes	-	DeepSeek-AI	Locally Load	General
DeepSeek-V3.2-Thinking (DeepSeek-AI et al., 2025)	Yes	-	DeepSeek-AI	Locally Load	General
DeepSeek-R1-0528 (Guo et al., 2025)	Yes	671B (37B active)	DeepSeek-AI	Locally Load	General

Table 11: Details of all evaluated LLMs.

- 北京：社会科学文献出版社，2017年。
- 关晓红：《科举停废与近代中国》，北京：社会科学文献出版社，2017年。
 - 韩策：《科举改制与最后的进士》，北京：社会科学文献出版社，2017年。
 - 李细珠：《张之洞与清末新政研究》，北京：中国社会科学出版社，2015年。
 - 王建朗、黄克武主编：《两岸新编中国近代史·晚清卷》，北京：社会科学文献出版社，2016年。
 - 罗志田：《权势转移：近代中国的思想、社会与学术》，武汉：湖北人民出版社，1999年。
 - 王先明：《近代绅士——一个封建阶层的命运》，天津：天津人民出版社，1997年。
 - 费孝通：《中国绅士》，北京：中国社会科学出版社，2006年。
 - 费孝通：《乡土重建》，长沙：岳麓书社，2012年。
 - 张希清、毛佩琦、李世愉主编；张希清著：《中国科举制度通史（宋代卷）》，上海：上海人民出版社，2015年。
 - 张希清、毛佩琦、李世愉主编；金滢坤著：《中国科举制度通史（唐代卷）》，上海：上海人民出版社，2015年。
 - 张希清、毛佩琦、李世愉主编；李世愉、胡平著：《中国科举制度通史（清代卷）》，上海：上海人民出版社，2015年。
 - 张希清、毛佩琦、李世愉主编；郭培贵著：《中国科举制度通史（明代卷）》，上海：上海人民出版社，2015年。
 - 张希清、毛佩琦、李世愉主编；武玉环等著：《中国科举制度通史（辽金元卷）》，上海：上海人民出版社，2015年。
 - 乔卫平著，李国钧、王炳照主编：《中国教育制度通史（宋辽金元）》，济南：山东教育出版社，2000年。
 - 刘海峰：《科举学导论（增订本）》，北京：中国社会科学出版社，2025年。
 - 刘海峰：《科举学十讲》，杭州：浙江古籍出版社，2025年。
 - 王日根等著，刘海峰主编：《中国科举通史（清代卷）》，北京：人民出版社，2020年。
 - 金滢坤著，刘海峰主编：《中国科举通史（唐代卷）》，北京：人民出版社，2020年。
 - 钱建状著，刘海峰主编：《中国科举通史（宋代卷）》，北京：人民出版社，2020年。

T1	T2	T3	T4
<p>Question: 对朱卷进行名词解释。Provide a definition for the term "Zhujuan (朱卷)".</p> <p>Reference Answer: 朱卷是科举制度中试卷明目之一。宋代真宗朝实行考卷誊录制，应试者试卷弥封后送誊录所由专门人员用红笔抄写，称为“朱卷”……“Zhujuan (literally "Vermilion Scrip") was a type of examination paper within the imperial examination system. During the reign of Emperor Zhenzong of the Song Dynasty, a system of transcribing exam scripts was implemented. After an examinee's paper was sealed to conceal their identity, it was sent to a transcription bureau where specialized scribes would copy it using red ink. This red-ink copy was called the "Zhujuan"...</p> <p>Rubrics: 加分项 Bonus Points: ...3. 考出设置朱卷的目的是避免考官徇私舞弊。Stating that the purpose of the Zhujuan system was to prevent favoritism and corruption among examiners. (+3) 4. 指出朱卷是誊录官员用红笔抄录考生墨卷形成的。Explaining that the Zhujuan was created by scribes who transcribed the candidate's "Mojuan" (Ink Scrip) using red ink. (+3)...</p> <p>减分项 Penalty Points: 1. 滥用学术表述，如生搬硬套社科理论、类比失当。Inappropriate use of academic jargon, such as forcing social science theories or making improper analogies. (-3)...</p>	<p>Question: 明中期之后“游戏八股文”依托着科举八股文出现，试从文体特征、审美取向及历史价值的角度介绍“游戏八股文”是什么。After the mid-Ming Dynasty, the 'Playful Eight-Legged Essay' emerged, building upon the official imperial examination's eight-legged essay. Please introduce what the 'Playful Eight-Legged Essay' is from the perspectives of its stylistic features, aesthetic orientation, and historical value.</p> <p>Reference Answer: ...在审美趣味上，游戏八股文展现出五对鲜明的矛盾特征：1. 俗与雅；题材和语言走向浅俗，追求玩世不恭的“俗趣”，借圣贤之酒杯浇胸中之块垒。... In terms of aesthetic taste, the Playful Eight-Legged Essay exhibits five pairs of distinct contradictory features: 1. Vulgarity vs. Elegance: Themes and language trended toward the vernacular, pursuing a cynical 'vulgar interest' and 'using the sages' wine cup to drown one's own sorrows.'...</p> <p>Rubrics: ...16. 答到游戏八股文按照表现形态区分，包括融入了八股要素的小说、戏曲、诗赋以及分咏体诗等多种交叉体裁。Listing a type classified by form: hybrid genres such as novels, dramas, poems, and 'fenyong-style shi zhong' that incorporated elements of the eight-legged essay. (+3) 17. 答到在审美趣味上，游戏八股文展现出五对鲜明的矛盾特征。Stating that in terms of aesthetic taste, the Playful Eight-Legged Essay exhibits five pairs of distinct contradictory features. (+2)</p>	<p>Question: 唐宋之间社会发生了巨大的变化，科举制既是这种变化的表现，也是这种变化的原因。针对科举制与唐宋间社会变化关系，学者们提出了哪些观点。Society underwent tremendous changes between the Tang and Song dynasties. The imperial examination system was both a manifestation and a cause of these changes. What viewpoints have scholars proposed regarding the relationship between the imperial examination system and the social changes between the Tang and Song dynasties?</p> <p>Reference Answer: ...20世纪初内藤湖南提出唐宋变革论，观点是：科举制度在唐宋间发生根本性转变。唐代科举仍具贵族性。至宋代，经王安石变革，改考经义（阐发经典义理）与策论（表述政见），选拔标准转向学识与政治见解。... In the early 20th century, Naitō Konan proposed the 'Tang-Song transition' theory. His viewpoint was that the imperial examination system underwent a fundamental transformation between the Tang and Song. The Tang examinations still had an aristocratic nature... By the Song dynasty, following the reforms of Wang Anshi, the focus shifted to examining 'classical exegesis' (explaining the meaning of the classics) and 'policy essays' (expressing political views)...</p> <p>Rubrics: ...3. 答出20世纪初内藤湖南提出唐宋变革论，观点是：科举制度在唐宋间发生根本性转变。Mentioning that in the early 20th century, Naitō Konan proposed the 'Tang-Song transition' theory, with the viewpoint that: the imperial examination system underwent a fundamental transformation between the Tang and Song. (+5)...</p>	<p>Question: 假设你是一位清朝道光癸巳科（道光十三年）的科举会试考生，请你严格按照八股文的写作方式写一篇科场八股文，如果遇到需要避讳的字，以拼音标注。字数不要超过700字。请直接作答，仅输出你写的八股文，不要输出其他内容。题目是：“古之愚也直，今之愚也佞而已矣。” Assume you are a candidate in the Metropolitan Examination of the Guisi year during the Daoguang era (the 13th year of Daoguang's reign) of the Qing Dynasty. Please write an examination-style eight-legged essay strictly following the required format. If you encounter any characters that are subject to naming taboos, mark them with Pinyin. The essay should not exceed 700 characters. Please provide your answer directly by outputting only the eight-legged essay, with no other content. The topic is: "In antiquity, stupidity was accompanied by straightforwardness; nowadays, stupidity is accompanied by deceitfulness."</p> <p>Reference Answer: 愚亦有古今之辨流弊与狂狷亦夫愚也而直乃成愚而佞者何也即古證今其流弊不與狂狷等哉... Stupidity, too, has its distinction between the ancient and the modern; its corrupting influence is equivalent to that of recklessness and arrogance. Now, why has straightforward stupidity become deceitful stupidity? Indeed, when we use the ancient to examine the present, are its corrupting influences not equivalent to those of recklessness and arrogance? ...</p> <p>Rubrics: 加分项 Bonus Points: 1. 文章包括：破题、承题、起讲、入题四个环节。The essay includes the four initial sections: Poti (Opening), Chengti (Elaboration), Qijiang (Preliminary Exposition), and Ruti (Initial Argument). (+5) ...5. 起讲要“入口气”，即不能用自己的口气来议论问题，而是用别人的身份和口吻议论。开头常用“且夫”“若曰”“以为”“崇思”等，也有单用“且”“而”的，均代表要代圣人讲话。The Qijiang must adopt an "impersonal tone" meaning it should not be argued from one's own perspective, but from the persona of another, such as a sage. It often begins with phrases like "Moreover", "If one were to say", "It is thought that", or "I have often considered". Single characters like "and" or "Furthermore" can also be used. All are meant to represent speaking on behalf of the Sage. (+3)...</p> <p>减分项 Penalty Points: ...3. 未对“玄”“辘”“胤”“正”“弘”“历”“颀”“臻”“璉”“璵”“璆”“中”“丘”以拼音标注避讳。Failure to mark the following taboo characters with Pinyin: "玄", "辘", "胤", "正", "弘", "历", "颀", "臻", "璉", "璵", "璆", "中", "丘". (-60)</p>

Figure 21: Sample questions, reference answers, rubrics from four tasks.

- 李兵著，刘海峰主编：《中国科举通史（辽金元卷）》，北京：人民出版社，2020年。
- 郭培贵著，刘海峰主编：《中国科举通史（明代卷）》，北京：人民出版社，2020年。
- 宫崎市定：《科举史》，郑州：大象出版社，2020年。
- 李成茂：《高丽朝鲜两朝的科举制度》，北京：北京大学出版社，1993年。
- 费正清、赖肖尔：《中国：传统与变革》，陈仲丹等译，南京：江苏人民出版社，2012年。
- Benjamin A. Elman: A Cultural History of Civil Examinations in Late Imperial China, Berkeley and Los Angeles: University of California Press, 2000.
- Henrietta Harrison: The Man Awakened from Dreams: One Man's Life in a North China Village, 1857–1942, Stanford: Stanford University Press, 2005.
- 安东强：《“中国政治史事论”与清末科举改制》，《文学遗产》，2021年第5期。
- 安东强：《论校注本〈钦定学政全书〉的文献价值》，《武汉大学学报（人文科学版）》，2012年第1期。
- 陈长文：《明代进士登科录的文献价值及其局限性》，《甘肃社会科学》，2006年第6期。
- 陈长文：《明代科举中的官年现象》，《史学月刊》，2006年第11期。
- 陈长文：《明代科举中的“告殿”现象》，《图书馆杂志》，2008年第4期。
- 陈胜：《一项备受争议的教育制度——清末学堂奖励出身制度述评》，《华东师范大学学报（教育科学版）》，2011年第1期。
- 陈文：《试析法国人对越南科举考试的影响》，载刘海峰、朱华山主编：《科举学的拓展与深化》，武汉：华中师范大学出版社，2013年。
- 陈维昭：《论清代科场条例修订与八股文体演变的关系》，《文艺理论研究》，2025年第4期。
- 陈宝良：《明代学官制度探析》，《社会科学辑刊》，1994年第3期。

Top-tier Academic Papers includes:

- 丁修真：《举人的路费：明代的科举、社会与国家》，《中国经济史研究》，2018年第1期。
- 龚延明、高明扬：《清代科举八股文的衡文标准》，《中国社会科学》，2005年第4期。
- 龚笃清：《试述明代前期八股文对文学的影响》，《中国文学研究》，2005年第1期。
- 郭培贵：《关于明代科举研究中几个流行观点的商榷》，《清华大学学报（哲学社会科学版）》，2009年第6期。
- 郭培贵：《明代庶吉士群体构成及其特点》，《历史研究》，2011年第6期。
- 郭培贵：《明代武举的形成与确立》，《明史研究》，2017年。
- 郭培贵：《明代科举中的座主、门生关系及其政治影响》，《中国史研究》，2012年第4期。
- 郭培贵：《明代会试分卷录取制创立、实施及变迁考实》，《史学集刊》，2024年第6期。
- 郭培贵：《试论明代提学制度的发展》，《文献》，1997年第4期。
- 高福顺：《辽朝礼部贡院与知贡举考论》，《考试研究》，2011年第2期。
- 关晓红：《晚清议改科举新探》，《史学月刊》，2007年第10期。
- 关晓红：《清代朝考之创制与终结》，《学术研究》，2016年第11期。
- 韩策：《科举改制与诏开进士馆的缘起》，《近代史研究》，2015年第1期。
- 黄强：《游戏八股文的文学趣味——介绍俗文学的一个新品种》，《江南大学学报（人文社会科学版）》，2011年第1期。
- 贾安琪、聂鑫：《清代〈科场条例〉纂修考述》，《浙江大学学报（人文社会科学版）》，2024年第11期。
- 姜新：《评清末民初的留学生归国考试》，《史学月刊》，2005年第12期。
- 李发根：《科举制的废除与近代中国乡村危机研究》，《山西师范大学学报（哲学社会科学版）》，2016年第6期。
- 李建军：《明代武举制度述略》，《南开学报》，1997年第3期。
- 李思成：《明清会试“阅本经”规则的演变》，《山东社会科学》，2023年第9期。
- 李思成：《身不由己：明代的座主门生与党争再探》，《山东社会科学》，2025年第1期。
- 李永明：《京师大学堂历史学科的发生与发展》，《史学史研究》，2025年第3期。
- 林浩彬：《裁撤府、州、县学教职与清末新政》，《社会科学战线》，2025年第9期。
- 刘海峰：《“科举”含义与科举制的起始年份》，《厦门大学学报（哲学社会科学版）》，2008年第5期。
- 刘海峰、毛鹏程：《清乾隆朝科举改革：动因、举措及影响》，《厦门大学学报（哲学社会科学版）》，2023年第2期。
- 刘海峰：《论述东亚科举文化圈的形成与演变》，《厦门大学学报（哲学社会科学版）》，2016年第5期。
- 刘海峰：《中国对日韩越三国科举的影响》，《学术月刊》，2006年第12期。
- 刘明鑫：《明代的科举走报》，《史学月刊》，2019年第7期。
- 刘明鑫：《明代会试考生路费资助制度考论》，《历史档案》，2020年第4期。
- 刘明鑫：《明代会试考生应考旅费考察》，《中国史研究》，2022年第4期。
- 刘晓琴：《严复与晚清留学生归国考试研究》，《南开学报》，2014年第1期。
- 刘希伟：《清代科举考试中的“商籍”考论——一种制度史的视野》，《清史研究》，2010年第3期。
- 刘志强：《越南阮朝科举及其本土特色》，《东南亚纵横》，2010年第4期。
- 兰婷、王成铭：《金代女真官学》，《社会科学战线》，2010年第9期。

- 毛晓阳:《清代宾兴礼考述》,《清史研究》,2007年第3期.
- 宋豪飞:《清代科举教科书:〈钦定四书文〉的编选与文风宗尚》,《中山大学学报(社会科学版)》,2025年第3期.
- 宋巧燕:《清代科举试帖诗写作规范探析》,《教育与考试》,2015年第3期.
- 申万里、窦相国:《元代儒户体制下士人的管理制度》,《社会科学研究》,2025年第5期.
- 薛瑞兆:《论金朝的“南北选”》,《学术交流》,2025年第1期.
- 郝志群:《封建科举、职官中的“官年”——从杨守敬的乡试卷谈起》,《历史研究》,2003年第4期.
- 萧启庆:《元代的儒户:儒士地位演进史上的一章》,载《元代史新探》,台北:新文丰出版公司,1983年.
- 萧启庆:《元代科举特色新论》,《台北“中研院”历史语言研究所集刊》,第81本,2010年.
- 王静:《清代科举中“岁科连考”的规制变通与实践效果》,《清史研究》,2025年第4期.
- 王庆成:《清代学政官制之变化》,《清史研究》,2008年第1期.
- 王日根、章广:《清代八旗科举制度的发展及其影响》,《考试研究》,2015年第5期.
- 王熹:《明代朝野对科举制度的评论》,《明史研究》,2001年.
- 王学深:《清代科举试卷违式问题探析》,《古代文明(中英文)》,2024年第4期.
- 王学深:《清代乾隆朝科举冒籍问题论述》,《中国考试》,2016年第4期.
- 王学深:《清代科举“未殿试”成因再探析》,《地域文化研究》,2022年第4期.
- 吴恩荣:《科考、遗才与大收:明代乡试资格考试述论》,《安徽大学学报(哲学社会科学版)》,2013年第5期.
- 吴光辉:《科举考试与日本》,《东南学术》,2005年第4期.
- 王方:《科举制度与儒学在日本的早期传播》,《外文研究》,2004年第2期.
- 杨胜祥:《清代科举“补殿试”对进士名次的影响》,《清史研究》,2020年第2期.
- 于晓燕:《“义学”释义》,《贵州师范学院学报》,2014年第10期.
- 张仲民:《“科举”余绪:清末最后的“杂试”与“朝考”》,《史学月刊》,2024年第7期.
- 张晓波:《朝鲜王朝科举制度研究》,山东大学师范大学博士学位论文,2020年.
- 郑天挺:《清代考试的文字——八股文和试帖诗》,《故宫博物院院刊》,1982年第2期.
- 赵利峰:《清中后期广东闹姓考原》,《暨南史学》,2003年.
- 左玉河:《论清季学堂奖励出身制》,《近代史研究》,2008年第4期.