

参赛队员姓名: Ziyan Hua

中学: Shanghai American School

省份: 上海市

国家/地区: 中国

指导教师姓名: Zhi-Qi Cheng & 肖俊

指导教师单位: 卡耐基梅隆大学 & 浙江大学

论文题目: HRBench: Historical Reasoning Benchmarking
for Large Language Models

2024 S.-T. Yau High School Science Award
仅用于2024丘成桐中学科学论文公示

HRBENCH: HISTORICAL REASONING BENCHMARKING FOR LARGE LANGUAGE MODELS

Ziyan Hua

Shanghai American School, Shanghai, China

ziyanhua@outlook.com

ABSTRACT

In recent years, the rapid progress of artificial intelligence (AI) has stimulated interdisciplinary research and applications. For example, the concept of AI for science has the potential to shift the paradigm of scientific research. Among these developments, one of the most influential AI breakthroughs, large language models (LLMs), has demonstrated tremendous potential for expediting the progress of social science research is an area highly related to language, and has received relatively limited attention so far. In this paper, we concentrate on integrating LLMs into historical research, a significant branch of social science. Our goal is to enhance historical reasoning ability through AI empowerment. Specifically, our work focus on the challenging task of multiple-source multiple-hop reasoning in historical research. However, a major obstacle that we encountered was the lack of relevant datasets to evaluate and improve LLMs in academic field. To alleviate this gap, we employed LLMs to generate a dataset for benchmarking and training purposes, which is a key focus of our study. We found that historical reasoning presents unique challenges compared to conventional language and logical reasoning tasks. Constructing a comprehensive, challenging, and assessable dataset was no easy task. Through iterative experimentation, we developed a workflow-driven method for creating the historical reasoning dataset automatically, resulting in a set of approximately 1,800 multiple-select questions that require multiple-source multiple-hop historical reasoning to solve. The dataset was validated by human historians, which shows that it has considerable complexity and high accuracy. We also proposed a method called HR-Tuning to improve general LLM performance in historical reasoning by leveraging source interpretation data and source deduction data from this dataset (which avoids over-fitting to the pattern of our data generation process). Extensive experimentation showed that the improved LLM significantly outperformed the original versions of the LLMs. Our contributions include the insights into historical reasoning and benchmark design, the automatic method of creating a challenging historical multiple-source reasoning dataset, and an improved LLM using supervised fine-tuning (SFT) for enhanced historical reasoning capabilities. As far as we know, this is the first work that advances the application of LLMs in historical research, laying the groundwork for more refined AI-driven methodologies in the social sciences.

1 INTRODUCTION

There have been significant breakthroughs in AI over the past decade, especially with the emergence of large language models (LLMs) such as ChatGPT (Floridi & Chiriatti, 2020; Achiam et al., 2023), Claude, and Llama, which has allowed for the application of AI in more disciplines. In this context, AI has increasingly become a core tool driving advancements in various academic disciplines, profoundly impacting research in both the natural and social sciences. With the development of AI technologies, researchers are now able to process massive datasets much more efficiently, reveal hidden patterns, and propose innovative solutions to complex problems.

In natural sciences, the integration of AI has significantly transformed how research is conducted. For instance, AI is now being applied in drug candidate discovery, where AI-powered models analyze molecular structures to rapidly identify potential drug candidates (Jumper et al., 2021). These

technologies not only improve the efficiency of scientific discovery or invention but also open new paths for AI-driven automated experimentation and ideation. Simultaneously, the applications of AI in the social sciences have been growing. The social sciences, focusing on human behavior and social structures, benefit from AI, specifically in economics (Hansen & Coauthors, 2023), sociology (Manning et al., 2024), and political science (Liu et al., 2023). However, LLM empowered social science research has gained less attention so far, despite being much more highly focused on language, compared with the natural science disciplines.

Working in a crucial field within the social sciences, historical researchers are applying AI technology for academic exploration, driving innovation in several key areas. The working principles of history, its theory and artificial intelligence are interlinked. As early as 1976, Joseph Weizenbaum (Agassi, 1976) proposed that it is necessary for history to be involved in artificial intelligence, especially to preserve historical materials. History is the conscious reflection of turning “data, memories, evidential relics about the past, documents and relics” into history through language (Foot & Partner, 2012), whose working mode is similar to that of LLMs that absorb and synthesize language materials. As Joshua Sternfeld said (Sternfeld, 2023), collecting model training data is like collecting historical materials. Upon deeper analysis, AI learning resembles historical research even more.

An example is the use of AI in translating ancient languages such as Akkadian to English using neural machine translation models, which provides historians with tools to better understand ancient civilizations without exhaustive manual translations (Gutherz, 2023). Additionally, there has been considerable emphasis on employing AI to the study of ancient documents in various areas, including digitization, restoration, attribution, language analysis, text criticism, translation, and decryption (Sommerschild et al., 2023).

In addition, some researchers (Garcia & Weilbach, 2023) find out that the advent of powerful LLMs offers a novel approach to facilitate a new form of inquiry for historians and Humanities researchers when augmented with vector embeddings from specialized historical academic sources, including primary sources, expert-written secondary sources, and combinations of the two. Their findings demonstrate that LLMs can leverage their semantic reasoning capabilities to analyze extensive textual archives. However, there is little work on using LLMs to do deeper historical research, especially in the direction of historical reasoning, which is one of the most challenging tasks in historical research.

In reality, though, reasoning has emerged as one of the most exciting topics in the current wave of interest in LLMs. This interest is particularly reflected in recent studies that explore various methodologies for enhancing reasoning capabilities, especially on structured thought processes.

In historical research, reasoning also plays an important role. History teachers have mentioned that historical reasoning, including the processes of change, causes, consequences, similarities, and differences in historical phenomena and periods, could help students to give meaning to the past (Van Boxtel & Van Drie, 2018). Furthermore, reasoning is also a crucial methodological foundation in historical science. Researcher (Shi, 2022) has pointed out that according to the concept of historical-causal narrative, researchers can engage in theoretical dialogues, test hypotheses, and further explore causal patterns within logic and evidence, thereby fulfilling the purpose of “learning from history” in the field of historical political science.

However, current LLM reasoning and historical reasoning differ fundamentally in four key areas: 1) Knowledge Base: LLMs rely on general knowledge and common sense drawn from vast datasets, providing broad but non-contextualized answers. In contrast, historical reasoning requires specific contextual knowledge of time, place, and circumstance to interpret past events accurately. 2) Complexity: LLMs follow clear and straightforward logical chains, while historical analysis transcends one singular aspect, and often involves multi-dimensional analysis, incorporating societal, economic, and cultural perspectives for a more comprehensive understanding. 3) Result Format: LLMs offer explicit and clear right answers based on clear evidence. Historical reasoning, by nature, leads to uncertain conclusions or theses, acknowledging multiple interpretations due to the ambiguous nature of historical evidence. 4) Evidence Usage: LLMs use structured, reliable evidence, synthesizing large amounts of complete information, and often produce clear conclusions. In contrast, historical reasoning deals with incomplete, unclear, or even contradictory evidence, demanding critical judgment to interpret diverse sources.

From these aspects, it is clear that the usage of evidence is both crucial and particularly challenging in historical reasoning. The reason is that it requires not only specialized knowledge grounded in the

specific historical context and conclusions with multiple perspectives, but also the consideration of various social, political, and cultural factors. Therefore, in this paper, we decided to focus on using evidence for historical reasoning empowered by LLMs.

Specifically, the usage of evidence has three major difficulties. Firstly, contradicting evidence makes it difficult to figure out the accurate version of an event. Historians have to analyze multiple sources and deal with the disparities. Secondly, incomplete evidence is an obstacle as crucial details may be missing, forcing the usage of creative approaches to fill in the blanks. Thirdly, lack of linguistic clarity in ancient texts and documents makes them hard to interpret, adding uncertainty. From the perspective of the common optimization path of LLMs, overcoming these challenges requires fine-tuning and testing on benchmark datasets to enable the models to perform historical reasoning with evidence. However, in existing LLM research, benchmarks related to historical reasoning cannot be found. It is hard to evaluate the performance of these LLM models in handling historical reasoning tasks without a high-quality dataset. Therefore, we decided to create a benchmark dataset for this purpose, which can be also used to improve the reasoning capabilities of LLMs in historical research.

In the process of building a historical reasoning benchmark, we faced significant challenges in ensuring the typical requirements of a dataset on comprehensiveness, complexity, quantity and quantifiability. One potential approach is to generate historical questions and reasoning answers using enhanced LLMs and then verified and checked by human’s historians. However, this adds more complexity to the workflow, as it requires careful control of the model’s output to maintain the quality of the benchmark.

To address these, we proposed a set of designing principles and success criteria on two perspectives. First is the historical perspective, which emphasizes the necessity of complex logical reasoning supported by evidence. This includes the ability to analyze, manipulate, and differentiate between multiple historical sources, as well as support multiple-hop reasoning. Second is the benchmarking perspective, from a data science standpoint, the focus shifts to ensuring that the benchmark maintains a sufficient level of difficulty to challenge advanced models, while still producing quantifiable results and creating comprehensive answers.

Overall, our research offers the following three significant contributions:

- **Insights in Historical Reasoning and Benchmark Design:** By looking into the complexity of historical reasoning using LLMs, it opens up new frontiers for researchers to better understand and efficiently analyze historical phenomena. It also enlarge the possibilities of LLMs in the field of social science research and serves as a solid foundation for further in-depth study.
- **Challenging Benchmark and Extensive Evaluation:** It makes a significant contribution by creating and presenting a previously nonexistent challenging benchmark for historical reasoning supporting multiple-source inputs. The dataset not only provides questions and answers, but also the “reasoning chain”, i.e., the intermediate reasoning steps, when answering the questions. Through extensive evaluation, it rigorously tests and validates the reasoning capabilities of different models.
- **Improved Tuning Method for Historical Reasoning:** The research proposes a SFT-based tuning method using source interpretation data and source deduction data, which significantly improves LLMs’ historical reasoning capabilities.

We want to make it clear that the above mentioned LLMs based historical reasoning approach is not the same thing as quantitative history research (Furet, 1971), which refers to the use of data and statistical methods to analyze and interpret historical events. It aims to uncover patterns, trends, and causal relationships in historical development through numbers and data. In contrast, historical reasoning with LLMs focuses on leveraging the reasoning capabilities of LLMs to analyze complex historical contexts, make inferences, and generate insights grounded in historical evidence, addressing the ambiguity and interpretative nature of historical research, which transcends the data-driven focus of quantitative history.

The following sections of our paper are outlined as follows: We will first review related works in AI reasoning, LLM benchmarking, and data generation. Secondly, we will introduce our data generation process using LLMs, following the design principles we summarized. We will also introduce a tuning scheme for open-source LLMs to improve their capabilities on historical reasoning. And lastly, we

will present experimental evaluation results, comparing different models and methods, and conclude with an analysis of results, limitations, and future work.

2 RELATED WORKS

The work proposed in this paper has correlated with three research directions in AI area, including AI reasoning, AI reasoning benchmarking and data generation.

2.1 AI REASONING

Notable recent research on enhancing LLM reasoning capabilities includes Chain-of-Thought (CoT) (Wei et al., 2022), Graph-of-Thought (GoT) (Besta et al., 2024), and Tree-of-Thought (ToT) (Yao et al., 2024).

CoT prompts LLMs to generate intermediate reasoning steps. The simple strategy significantly improves the performance on tasks requiring multi-step reasoning, such as math problems and code generation problems. CoT also improves the interpretability of LLMs. Inspired by CoT, GoT further explores multiple reasoning paths in parallel. This enable the models to handle more complex and non-linear reasoning tasks, such as decision-making problems with multiple perspectives. ToT organizes intermediate reasoning steps into a tree structure. This allows the models to branch and prune solutions, making it effective for tasks requiring iterative refinement and uncertainty handling.

Although significant progress has been made in improving the reasoning capabilities of LLMs, current reasoning tasks are still limited to general reasoning problems, which do not examine and utilize a wide range of evidences (sources) to make arguments and draw deductions like the way in historical research.

2.2 AI REASONING BENCHMARKING

Benchmarks help evaluate the capabilities and the limitations of LLMs across various domains and tasks. Currently, there are two well-known benchmarks that are related to the AI reasoning tasks.

The first is the MMLU (Massive Multitask Language Understanding) dataset (Hendrycks et al., 2020), which is designed to evaluate LLMs on 57 tasks in areas as humanities, STEM, and social sciences. It features multiple-choice questions based on academic and professional exams, testing both factual knowledge and reasoning skills. MMLU is commonly used to assess models' abilities in zero-shot and few-shot settings, making it a key benchmark for understanding LLMs' generalization and reasoning capabilities. Although the MMLU involves historical questions, it only involves reasoning based on a single historical source and does not involve multi-step reasoning. Moreover, the question is not sufficiently difficult. Most existing LLMs get nearly full marks, lacking sufficient discrimination in historical reasoning capabilities. In addition, no historical reasoning chain is provided.

The second is HotpotQA dataset (Yang et al., 2018), which focuses on evaluating LLMs on multi-hop reasoning. It includes over 100,000 questions to challenge LLMs to differentiate relevant from irrelevant information across multiple documents, testing their abilities in comprehension and explanatory reasoning. Though it allows for answering questions based on two (but not more) documents, it is not specialized in the field of history, and no historical reasoning chain is provided.

In short, neither dataset effectively addresses the challenges of historical reasoning benchmarking.

2.3 DATA GENERATION

In recent LLM related research, both OceanGPT (Bi et al., 2023) and GeoGPT (Zhang et al., 2024) present advanced frameworks for domain-specific data generation, yet they follow different approaches tailored to their respective fields.

OceanGPT is designed for ocean-specific science tasks, in which the data generation process begins with a manually selected seed dataset built by human experts in ocean science. This initial data is then expanded through iterative process between the agents and a LLM such as GPT-3.5-turbo.

GeoGPT focuses on geo-spatial tasks, which enhances the efficiency of geo-spatial workflows by using LLMs' semantic understanding ability to automate the process of GIS data generation. By selecting and executing the necessary GIS tools, the system automatically adapts to different tasks with reduced human intervention.

Both methods emphasized the power of knowledge-enhanced data generation by expanding on seed datasets through LLMs. However, neither of these approaches explicitly considered reasoning capabilities in the data generation process. Instead, they primarily focus on enriching knowledge and overlook the reasoning process that would allow the models to handle more complex data. This limitation indicates that though these methods are successful in specific domains, they lack of the abilities of evaluating and improving the reasoning capabilities of LLMs.

2.4 SUMMARY

In summary, our paper will be quite different from existing work in three key directions. First, unlike general AI reasoning research, we focus on historical reasoning, emphasizing the use of diverse and multiple sources to build structured arguments. Second, existing benchmarks like MMLU and HotpotQA either lack complexity or are not specialized in history, while we aim to create a benchmark centered on multi-source, historically-informed reasoning. Finally, unlike OceanGPT and GeoGPT, which focus on domain knowledge expansion, we explicitly integrate reasoning into our data generation process, combining knowledge with intermediate steps for complex historical tasks.

We want to emphasize why it is necessary to generate a historical reasoning chain in the reasoning questions. Firstly, it enforces the model to be more reliable when generating correct answers, as well as provides stronger supervision for model tuning. Secondly, it is more convenient for human historians to verify the correctness of the generated data.

3 DATASET GENERATION

In this section, we introduce the historical reasoning benchmark design and propose an automatic method for constructing a challenging historical reasoning benchmark dataset. Without loss of generality, this method is based on Chinese historical reasoning since the high-quality Chinese historical research papers are currently more available.

3.1 BENCHMARK DESIGN

Benchmark Task Format. We chose **multiple-select questions** as the task format for our historical reasoning benchmark. There are four main characteristics of our multiple-select questions, as shown in Figure 1. First, each question have four options. Second, a single question could have multiple correct options. Third, each of our questions contains a list of historical sources (could be more than 10 sources). Fourth, in addition to the correct answers, we propose a novel **historical reasoning chain** design that equip each option with a reasoning chain that explain why this option is true (correct) or false (incorrect) according to the given sources. A chain could have multiple steps. Here is an example of a 4-step historical reasoning chain design:

Example: Explanation1 (Source1) + Explanation2 (Source2) -> Deduction

1. 李继迁将战略目标指向灵州, 宋军在西北陷入被动 (史料1) + 宋军辎重部队在瀚海遭党项围歼, 知灵州李守恩等边防要员阵亡 (史料8) -> 灵州成为双方争夺的重要战略要地 (推论1)
2. 张齐贤主张联合凉州蕃部援助灵州, 并因此与主弃派产生矛盾 (史料5) + 李沆和张齐贤在是否守灵州的问题上有重大分歧 (史料2) -> 朝廷内部存在主守派和主弃派的分歧 (推论2)
3. 清远军失陷后, 知枢密院事周莹斥责主弃派 (史料5) + 党项部袭击灵州引发朝野震动 (史料8) -> 灵州的守弃问题引发朝廷主要派系之间的政治博弈 (推论3)
4. 推论1 + 推论2 + 推论3 -> 灵州的守弃之争是朝廷内主要派系 (主守派和主弃派) 之间政治博弈的具体体现。

As shown in the historical reasoning chain above, each step (row) of the chain have a reasoning operator (denoted by “->”). To the right of the operator is a historical deduction, and to the left of the operator is at least two explanations of the given sources or previously-generated historical deductions. Notably, if the option is true, the deduction in the final row should explicitly prove the option is true. Otherwise, the final deduction should explicitly prove the option is false.

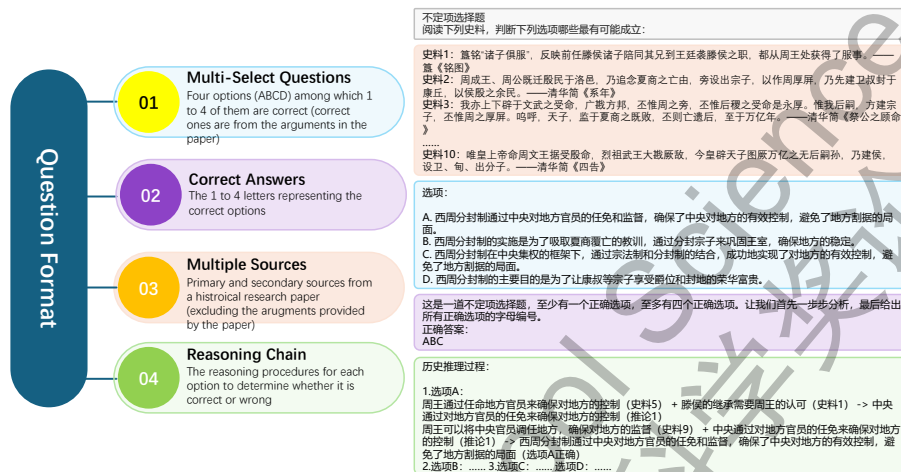


Figure 1: Benchmark Question/Task Format

Seed Data. To generate high-quality questions, we selected 1,781 publications from top Chinese journals in the field of ancient Chinese history as our seed data. All these papers are published between 2021 and 2024. These papers form a reliable foundation of our benchmark since they contain multiple historical sources and inherently present well-reasoned viewpoints by reasoning over these sources. Thus, the well reasoned viewpoints can be taken as the true options, the relevant sources are taken as the input sources and the underlying reasoning logistics constitute the reasoning chains. The false options, on the other hand, are generated based on the aggregated sources of all true options along with their own counter-argument chains.

LLM-based Data Generation Workflow. To further improve the quality of our generated questions, we propose a novel workflow-based data generation method for constructing historical reasoning data. We attempted all-in-one generation method, but discovered that it is excessively challenging. None of existing LLMs is capable of consistently producing high-quality benchmark data. In contrast, the workflow-based generation decomposes the challenging task, enabling more reliable data generation.

Since LLMs may still produce unreliable results even when we employ work-flow based method, we add a step to check the outputs one by one using commercial LLMs. We examine each option and its corresponding reasoning chains using the base LLM (GPT-4o). If the check failed, we may run the generation process multiple times. More details are in Section 3.2.

3.2 LLM-BASED DATA GENERATION WORKFLOW

This subsection introduces the data generation workflow of our historical reasoning benchmark. We use GPT-4o as the base LLM for the workflow. Figure 2 presents the five main steps of the workflow.

Step 1: Reading and Parsing the Paper Content

Procedure 1: Paper pre-processing

Input: Seed paper PDF

Output: Seed paper text

Step 1.1 Collecting historical research papers;

Step 1.2 Convert PDF into markdown file by MinerU;

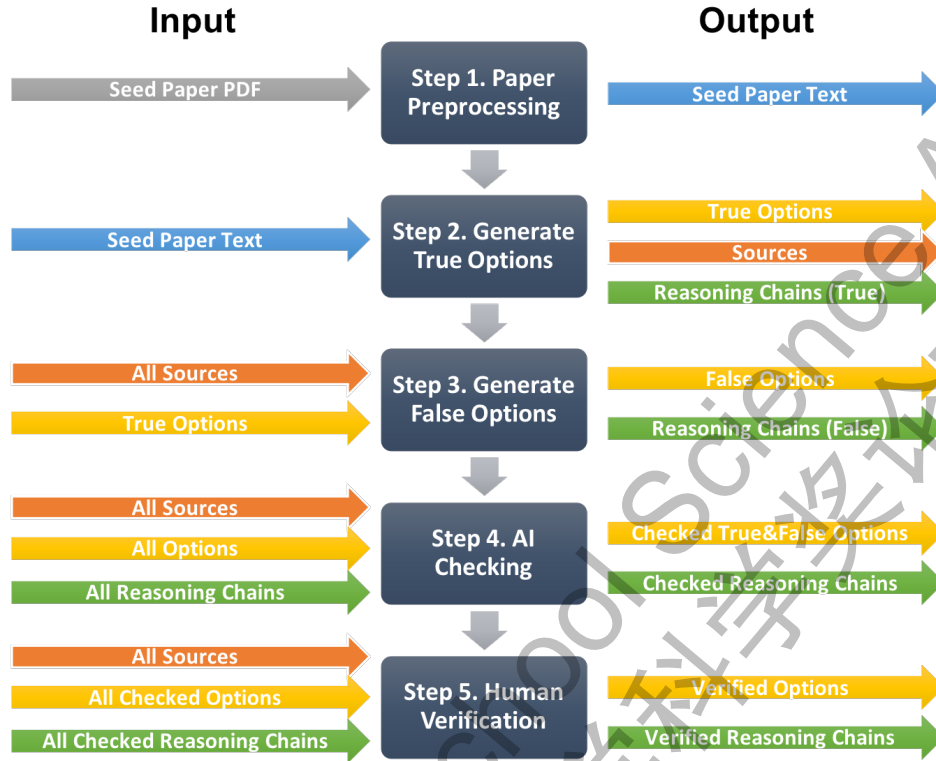


Figure 2: The workflow of our LLM based Generation

As shown in Procedure 1, the first step in our workflow involves reading and parsing the content of historical research papers. We employ the open-source document parsing tool, MinerU (Contributors, 2024), to convert PDF documents into machine-readable formats (markdown). MinerU excels in parsing complex layouts and converting symbols, making it particularly suitable for extracting text from scientific and historical literature. This tool enables us to accurately extract textual content, which forms the basis for subsequent steps in our workflow.

Step 2: Generating True Option Candidates

Procedure 2: Generate true options

Input: Seed paper text

Output: True options, sources, and reasoning chains of true options

Step 2.1 Extracting author’s viewpoints as true options;

Step 2.2 Listing relevant historical sources;

Step 2.3 Generating the historical reasoning chains;

As shown in Procedure 2, given the text of seed paper as input, this step generates true options, the sources relevant to these true options, and the corresponding reasoning chains.

Step 2.1: Extracting author’s viewpoints as true options. Given the parsed historical paper, we extract the author’s viewpoints as the true options. According to the common practice in historical reasoning, each viewpoint is supported by at least 2 historical sources.

Step 2.2: Extracting relevant historical sources. We list all historical sources that support the extracted viewpoint in Step 2.1. These sources might include primary and secondary sources.

Step 2.3: Constructing the historical reasoning chains. We extracted the underlying historical reasoning process of the viewpoint as a structured historical reasoning chain. This chain ensures that each step of the reasoning is grounded in historical sources and adheres to both historical common sense and basic logic. Each step in the reasoning chain must be supported by at least two inputs,

either historical sources or previously-generated deductions. The final output (deduction) of the reasoning chain should align with the extracted viewpoint in Step 2.1.

Step 3: Generating False Option Candidates

Procedure 3: Generate false options

Input: All sources, and true options

Output: False options, and reasoning chains of false options

Step 3.1 Creating misleading viewpoints;

Step 3.2 Constructing counter-arguments;

As shown in Procedure 3, given all previously-extracted sources and the true options as input, this step generates the false options and their reasoning chains.

Step 3.1: Generating distractor viewpoints. Based on the provided historical sources and existing viewpoints, we first generate distractor viewpoints that could be easily misinterpreted by a layperson. These viewpoints should be sufficiently complex and require at least 2 historical sources to discern their inaccuracy. It is discouraged to use absolute statements since this might degrade the difficulty levels of the questions.

Step 3.2: Constructing counter-argument reasoning chains. We also construct historical reasoning chains for the generated false options. The chains are similar to those of the true options, but with the final step refuting the input false options.

It is important that the incorrect viewpoints do not overlap with existing viewpoints. Thus, we use the word similarity measure to ensure the non-redundancy of the generated false options.

Step 4: Verifying Generated Options with AI

Procedure 4: AI checking

Input: All Sources, All Options, All Reasoning Chains

Output: Checked True&False Options, Checked Reasoning Chains

Step 4.1 Checking true option candidates;

Step 4.2 Checking false option candidates;

Step 4.3 Generating the final question;

As shown in Procedure 4, given all of the options, the sources and the reasoning chains we extracted/generated in previous steps as input, we generate the checked true&false options, and checked reasoning chains.

Step 4.1: Checking true option candidates. For each true option, this step verifies the correctness of the option and its corresponding reasoning chain.

1. Analyzing the option: confirming the accuracy of the option and ensuring that each step in the reasoning chains adheres to reasoning constraints.
2. Modifying the option: if the option is incorrect, we make minimal necessary modifications to correct it.
3. Modifying the reasoning chain: if the reasoning chain is unsatisfactory, we make necessary adjustments to ensure that each step is supported by at least two historical sources or previously-generated deductions, ultimately supporting the modified option.

Step 4.2: Checking false option candidates. For each false option, this step checks its correctness, its difficulty level and its own counter-argument chain.

1. Checking the false option: if the option is correct or lacks sufficient difficulty, we modify it to ensure it is incorrect and challenging to discern.
2. Checking the reasoning chain: if the counter-argument reasoning chain is unsatisfactory, we modify it to ensure that each step is supported by at least two historical sources or previously-generated deductions, ultimately refuting the modified option.

Step 4.3: Generating the final question. We generate the final multiple-select question by randomly selecting one to four true options from all true option candidates with three to zero randomly-selected false options, ensuring that there are a total of four options. The selected options, along with their corresponding historical sources, answers, and reasoning chains, are concatenated to form a complete multiple-select question. This format ensures clarity, transparency, and reliability, making it suitable for both human and LLM evaluation.

Step 5: Verifying Generated Options with Human Experts

Finally, we invited historical experts to review both the answers and the reasoning chains to ensure they accurately provide appropriate historical reasoning challenges and present reliable historical reasoning chains. We emphasize that even as more advanced AI algorithms emerge in the future, the “historian-in-the-loop” approach remains essential. This approach is crucial for ensuring the academic rigor of research, minimizing bias, incorporating scarce historical materials, and enhancing transparency. For detailed scoring & verification instructions for human’s historical experts, please refer to Appendix A.3.

3.3 THE PROMPTING FOR DATA GENERATION

Complex prompting is applied in the step 2, 3 and 4 as aforementioned. We have tested several commercial LLMs and found that GPT-4o demonstrated the best instruction-following ability. Moreover, we discovered that structured prompting yielded much better results compared with unstructured prompting. Thus, we designed specific structures for each generation or verification task, including several general components such as Goals, Knowledge, Constraints, Input Format, Output Format, and Examples. One thing to be pointed out is that using examples is crucial according to our experiments. We provide at least three well-selected examples for each step. In addition, experience also shows that repeating the the most important constraints (e.g., formats of reasoning chains) at different locations of the prompt gives better results. Details of the prompting structure of each step can be found in Figure 3.

3.4 SUMMARY

The detailed workflow ensures that our historical reasoning benchmark questions are both comprehensive and challenging, reflecting the complexities of historical reasoning and source understanding. By allocating specific evidence pieces to each actual answer option and adhering to a specified formatting of explanations (i.e., chains), we enhance the precision and the interpretability of our benchmark, making it a robust tool for evaluating historical reasoning skills. The code and the prompt of the above mentioned steps could be found in Appendix A.1 & A.2. More specific generated examples can be found in Appendix A.4 to gain a more intuitive understanding of the format we have set.

Figure 4 presents the statistics of the historical reasoning dataset (HRBench). It contains 1781 multiple-select questions generated from 1781 historical research papers. The numbers of correct options are nearly evenly distributed. Most questions have seven to twelve sources and 91% of them have at least seven sources, which demonstrates the complexity of the input information for the reasoning. Ninety percent of the questions have more than one reasoning step, and 77% of them have at least three reasoning steps, which shows the difficulty of multiple-hop reasoning.

4 HISTORICAL REASONING TUNING (HR-TUNING) FOR LARGE LANGUAGE MODELS

In this section, we will introduce our fine-tuning method on the open-source LLMs to improve its performance on historical reasoning.

4.1 ADVANTAGES OF FINE-TUNING LLMs FOR HISTORICAL REASONING

Unlike prompt engineering, which relies on contextually crafted prompts for each specific task, we use semi-supervised fine-tuning (SFT) due to the following reasons:

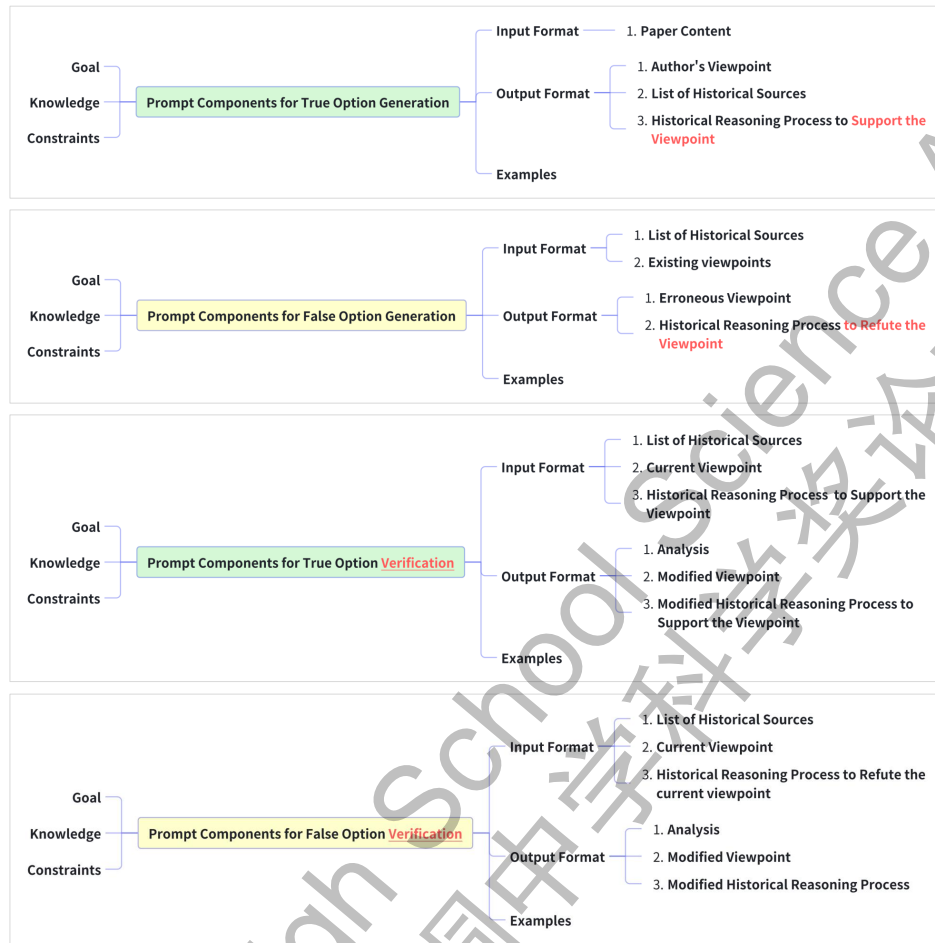


Figure 3: Prompting Structures

In total 1781 questions generated from 1781 historical research papers

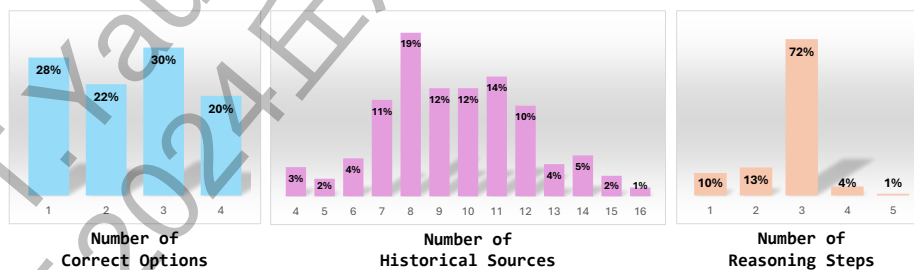


Figure 4: HRBench Data Statistics

1. Fine-tuned model could be better at understanding historical materials written in both classical Chinese and vernacular styles. This ensures that the model develops a nuanced understanding of the linguistic diversity inherent in historical documents.
2. It facilitates the model's comprehension of the political, economic, cultural, and social contexts specific to different historical periods. These contexts often require cross disciplinary knowledge that prompt engineering cannot achieve consistently across various temporal and spatial information.

3. The computational cost is reduced by relying on the embedding knowledge in the model, which makes it more efficient compared to the prompt engineering.
4. Since we have already utilized GPT-4o with prompt engineering to generate the HRBench, it might not get further improved by still using the prompt engineering on open-source models.

4.2 HISTORICAL REASONING TUNING (HR-TUNING)

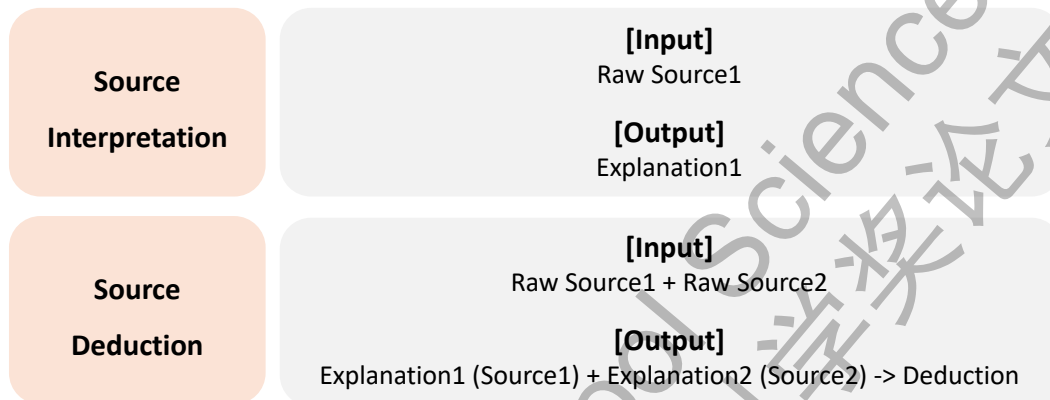


Figure 5: The illustrations of two kinds of historical reasoning tuning tasks in HR-tuning: the source interpretation tuning (the upper half) and the source deduction tuning (the lower half)

To improve the historical reasoning capabilities of existing LLMs, we introduce historical reasoning tuning (HR-tuning). HR-tuning contains two novel instruction tuning tasks: the source interpretation tuning and the source deduction tuning. By fine-tuning open-source LLMs, we empower existing LLMs with deeper understanding of historical sources and stronger multiple-source historical reasoning capabilities.

The source interpretation tuning. In the upper half of Figure 5, the model is provided with one primary or secondary historical source and is required to interpret the content, aiming to extract meaningful insights or implications from the text. The goal of this tuning method is to enable the model to capture the important information in the individual sources.

The source deduction tuning. As shown in the lower half of Figure 5, we provided multiple primary/secondary historical sources to the model and ask it to interpret the sources and draw coherent and appropriate deduction based on the sources. In the historical research area, the multi-hop reasoning process could face ambiguous, incomplete, and even contradictory evidence, which is extremely challenging.

The training data for both tuning tasks could be derive from the reasoning chains in the created HRBench dataset. However, to avoid over-fitting to the pattern of the data generation process (e.g., memorizing the styles of true/false options), we use only the reasoning chains that do not contain the final true or false options. They are typically the early reasoning steps of historical reasoning since the last step must contain true or false options according to our data generation procedures.

These two instruction tuning mimic the complexities of real-world historical reasoning and pose challenging SFT tasks for LLMs. More details are in Section 5.2 and Appendix A.1.2 and A.2.2.

5 EXPERIMENTS

In the following, we first evaluate the quality of the proposed Historical Reasoning Benchmark (HRBench). Then, using the up-to-date split of HRBench, we benchmark the historical reasoning capabilities of state-of-the-art (SOTA) open-source LLMs and closed-source LLMs. At last, we improve Qwen2-7B (Yang et al., 2024) with the proposed HR-tuning method, compare the resulting model (i.e., HR-Qwen2-7B) to the SOTAs and ablate its major design specifics.

5.1 MEASURING THE QUALITY OF HRBENCH

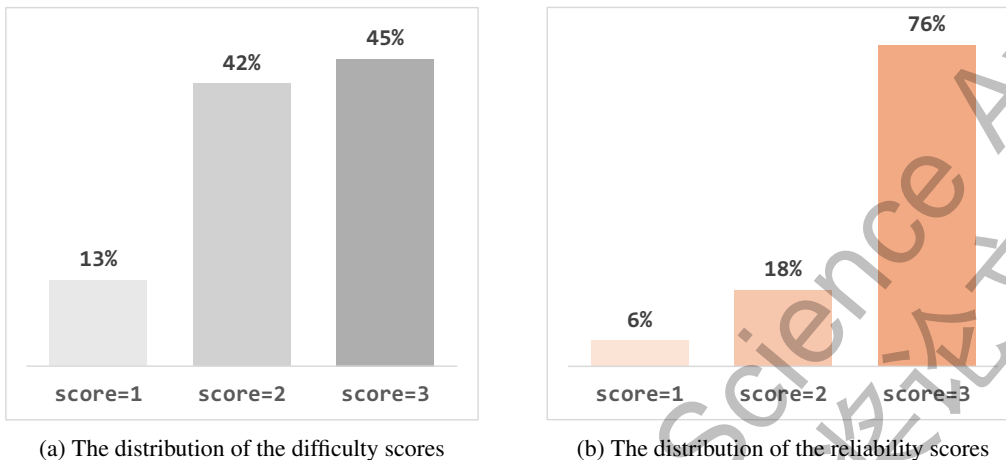


Figure 6: The distributions of the difficulty/reliability scores evaluated by expert historians

We use the up-to-date split of HRBench (namely HRBench-2024) as the benchmark dataset in this work. It contains 299 multiple-select questions and each of them is generated according to a historical paper published in 2024. A few expert historians were instructed to evaluate and verify of the proposed benchmark. First, the experts are required to score the questions from two perspectives: the difficulty level of the questions and the reliability level of the historical reasoning chains. The scores range from 1 to 3, with 1 being the lowest and 3 being the highest. The distributed are shown in Figure 6. On the one hand, the average difficulty score is 2.32 and approx. 42%/45% of questions are regarded as medium/high difficulty level by human experts. This verifies that HRBench is a challenging historical reasoning benchmark from the human perspective. On the other hand, the average reliability score is 2.69 and approx. 76% cases of historical reasoning are thought high-quality by human experts. This demonstrate the quality of the proposed benchmark. Second, following the guidelines detailed in Appendix A.3, all questions are further verified by the expert historians, which further improves the quality of the proposed benchmark.

5.2 EXPERIMENTAL SETUP

Benchmarking state-of-the-art (SOTA) LLMs. To explore the historical reasoning capabilities of existing LLMs, we perform a comprehensive evaluation on both closed-source LLMs and open-source LLMs. The closed-source LLMs are: GPT-4o (the 2024-05-13 version), Qwen-Max (the 2024-04-28 version), DeepSeek-V2.5 and Doubao-pro-32k (the 2024-08-28 version). The open-source LLMs are: Qwen2-7B (Yang et al., 2024), LLaMA-3-8B (Dubey et al., 2024), DeepSeek-Coder-6.7B (Guo et al., 2024) and GLM-4-9B (GLM et al., 2024). For all these models, we use their “chat/instruct” versions and set the temperature to 0. We assess the LLMs using the aforementioned HRBench-2024 (299 questions) and its hard subset HRBench-2024-Hard (136 questions with the difficulty scores as 3). Every question may have multiple true options, the model is judged to have answered correctly only if it identifies all correct and incorrect options. We utilize the metric Accuracy (%) to measure the model’s performance.

Historical Reasoning Tuning (HR-Tuning) for LLMs.

We uses historical reasoning chains to turn the model. To avoid over-fitting to the pattern of the data generation process (e.g., memorizing the styles of the options), we use only the 2021-2023 splits from HRBench (i.e., excluding the 299 human verified questions from the 1781 questions) for training and remove all rows that contains the true or false options (i.e., the final row of the chains). The 299 questions in HRBench-2024 are applied to test the performance of LLMs, and the part with difficulty score of 3 forms a subset called HRBench-2024-Hard. For all experiments, we adopt the Qwen2-7B model with full-parameter fine-tuning for 600 steps using an AdamW optimizer (Loshchilov et al., 2017) (with hyper-parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.95$ and $\text{weight_decay} = 0.1$). We use a global batch size of 16 and a cosine learning rate schedule with initial learning rate 2×10^{-6} . The gradient

clipping norm is 1.0 and the maximum sequence length is 2048. We conduct all experiments on 4 NVIDIA-A800 with ZeRO-3 data parallelism (Rajbhandari et al., 2020) and the LLaMA-Factory codebase (Zheng et al., 2024).

5.3 BENCHMARKING STATE-OF-THE-ART LLMs

Table 1: The results of state-of-the-art (SOTA) LLMs on HRBench-2024 and HRBench-2024-Hard. **Bold** denotes the best performing model on the dataset

LLMs	Accuracy on HRBench-2024	Accuracy on HRBench-2024-Hard
Qwen-Max (2024-04-28)	60.54%	60.29%
GPT-4o (2024-05-13)	60.20%	58.82%
Doubao-pro-32k (2024-08-28)	57.53%	57.35%
DeepSeek-V2.5	56.52%	50.74%
Qwen2-7B Yang et al. (2024)	56.52%	54.41%
GLM-4-9B GLM et al. (2024)	52.17%	47.06%
LLaMA-3-8B Dubey et al. (2024)	21.74%	19.12%
DeepSeek-Coder-6.7B Guo et al. (2024)	4.35%	2.21%
HR-Qwen2-7B (Ours)	65.55%	63.97%

Table 1 list the historical reasoning results of both open-source LLMs and closed-source LLMs on HRBench-2024 and HRBench-2024-Hard. The results show that Qwen-Max emerges as the highest-performing closed-source model with accuracy of 60.54%/60.29%, while GPT-4o, which also serves as the question generator, achieves comparable results of 60.20%/58.82%. While being much smaller than the closed-source LLMs, the performance of Qwen2-7B is highly competitive (56.52%/54.41%), and significantly outperforms other open-source baselines like GLM-4-9B, LLaMA-3-8B, and DeepSeek-Coder-6.7B. This is probably because the LLMs have difficulty accessing high-quality historical reasoning data during their pre-training phases, then their historical reasoning capabilities do not effectively grow with an increase in model size. Finally, the proposed HR-tuning method (HR-Qwen2-7B) achieves the best performance of 65.55%/63.97%, surpassing the second-best model (Qwen-max) by 5.01%/3.68% and the baseline model (Qwen2-7B) by 9.03%/9.56%. This verifies the effectiveness of HR-tuning in improving the historical reasoning capability of LLMs.

5.4 ABLATION STUDY ON HR-TUNING

Table 2: The ablation results of HR-tuning on HRBench-2024 and HRBench-2024-Hard. **Bold** denotes the best performing model on the dataset

LLMs	Accuracy on HRBench-2024	Accuracy on HRBench-2024-Hard
<i>a0</i> Qwen2-7B (Baseline)	56.52	54.41%
<i>a1</i> + source interpretation	60.20%	56.62%
<i>a2</i> + source deduction	65.55%	63.97%
<i>a3</i> + source deduction (w/o references)	62.21%	58.82%
<i>a4</i> + source interpretation & deduction	65.22%	62.50%

This subsection ablates major design elements in HR-tuning. All ablation experiments are based on Qwen2-7B (the “chat” version). (Yang et al., 2024).

Table 2 shows the ablation results of HR-tuning. First, compared to the baseline, introducing the source interpretation tuning alone improves the accuracy by 3.68% (*a1* vs. *a0*), matching the accuracy of GPT-4o on HRBench-2024 at 60.20%. We attribute this performance gain to the improved capabilities of LLMs in understanding classical Chinese. In comparison, introducing the source deduction alone significantly boosts the performance by 9.03% (*a2* vs. *a0*). This indicates the

effectiveness of training the model to deduce information from multiple sources, a critical capability in historical reasoning tasks where conflicting narratives are common. However, removing the references (explicitly linking the explanations to their sources) in the deduction output degrades the accuracy by 3.31% ($a3$ vs. $a2$). This phenomenon reveals the explicit deduction over multiple historical sources is crucial to the historical reasoning task. By combining both the source interpretation tuning and the source deduction tuning ($a4$), surprisingly we observe that the combination compromises the performance compared to $a2$ (65.22% vs. 65.55%). This may be because the interpretation data is inherently contained in the deduction data. During training, the model may resort to shortcuts by memorizing previously seen interpretation, instead of truly understanding the source in a multi-source context as in the deduction-only setting. This degrades the effectiveness of source deduction.

6 CONCLUSION

This paper focuses on benchmarking and improving the historical reasoning capabilities of large language models (LLMs). To resolve the unique challenges of historical reasoning, we construct a comprehensive, challenging, and assessable dataset HRBench, containing around 1,800 multiple-select questions that require strong multiple-source historical reasoning capabilities to solve. The dataset is validated by human historians and shows considerable complexity and high accuracy. We also proposed a method called HR-Tuning to improve LLMs' historical reasoning capabilities by leveraging source interpretation data and source deduction data from the HRBench dataset, aiming at avoiding over-fitting to the pattern of the proposed data generation process. Extensive experimentation showed that the improved LLM (HR-Qwen2-7B) significantly outperformed its original model and surpassed the state-of-the-art commercial models (e.g., GPT-4o and Qwen-max). We hope that this work can promote AI-driven methodologies in social science research.

7 REMARKS

7.1 LIMITATIONS AND FUTURE WORKS

While our research has made significant contributions to historical reasoning using LLMs, several limitations should be considered. First, the creation of the historical reasoning benchmark dataset is constrained by the availability and the quality of seed historical papers, while the reliance on several human experts for validation introduces potential biases and limits scalability. However, this limitation can be simply fixed by achieving more extensive cooperating in historical research community. Second, the fine-tuning process uses only 7B LLMs due to the limitation of the computational resources. When more resources are allocated, this disadvantage could be easily fixed. Third, the experimental evaluation is limited by the scope of the validation dataset and the metrics used. Reliance on accuracy as the primary metric may not fully capture the nuanced aspects of historical reasoning. To solve this, we intend to evaluate LLMs' capabilities of generating reliable historical reasoning chains in our future works. Fourth, the applicability of our findings to broader historical contexts and different languages remains uncertain. We plan to extend our method with more language by collecting high-quality seed papers in more languages and adopt multi-lingual LLMs.

In addition, our ultimate goal is to build a platform that assists historians in conducting their research more efficiently and effectively. Besides reasoning, historical studies involve a range of methodologies or tasks, such as comparative study, cliometrics, documentary analysis, and psychological history, among others. While it is possible to train a huge LLM using standard method to manage all these tasks, creating specifically designed datasets to enhance those capabilities could prove to be more effective.

To reach this ultimate goal, alongside the methodologies employed in historical studies, it is also vital to find an effective way to merge the strengths of LLMs with human expertise and intelligence. While some current applications do incorporate human interaction, like through prompt engineering, most LLM-based technologies primarily depend on the models' own capabilities. Historically, both historical and broader social science research have heavily relied on professional expertise. An LLM-powered research platform provides a new avenue for improving the efficiency of research tasks. However, sufficiently integrating human expert capabilities into the interaction with LLMs poses a

unique challenge compared to recent advancements seen in academia and industry. Essentially, this type of research platform not only depends on the LLM’s capabilities but also effectively harnesses the research capabilities of human experts. The quality of the outcomes will depend significantly on the varying levels of expertise among the human researchers involved.

The final step in historical research tasks involves composing the research report, a component that is currently also lacking. In this phase, the researcher provides guidance to the large-scale model, instructing it to integrate all preceding discussions and relevant literature into a coherent research paper, complete with references listed at the end of the document. Alternatively, the researcher might first draft an outline of the paper based on earlier discussions, following which the large-scale model can compile all available information into a well-structured academic paper.

While this is no easy task, we believe that achieving this ultimate goal is ultimately possible. Such a platform will not only enhance historical research but also prove valuable for other social science studies. Furthermore, it may offer potential insights for natural science research, although realizing this may present even greater challenges.

7.2 ONLINE DEMOS

Please visit our online demos through <http://www.hrllm.online>. A chat-bot interface is provided in the demo, where users can interact with the bot with any historical questions, though it is more focused on improving the LLM capabilities in multiple-source reasoning. For the convenience of trying the bot, we also provided a list of sample reasoning questions with multiple sources next to the input box. You can click any of the questions to try the improved historical LLM.

In addition, we also provide a link to test the multiple-select question generation process. You may upload a historical research paper in Chinese language and in PDF format (please be noted a scanned PDF will not work currently).

ACKNOWLEDGMENTS

I’d like to thank both Dr. Cheng and Prof. Xiao for their helpful and insightful comments on my topic selection process and paper writing process, in which they’ve significantly helped to improve my understanding of scientific research practices, skillful communication, and accurate representations of data. I’d also like to thank Dr. Cheng for his directing and instructing throughout the entire research process that helped me to push forward despite significant adversity. I’d like to thank the digital age for allowing me to access so much high-quality information and open-source AI models/tools at my fingertips to empower myself to conduct this research. Finally, I’d like to acknowledge my parents’ unwavering support for my research. Without all of this support, this research wouldn’t have been possible.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Joseph Agassi. Computer power and human reason: From judgment to calculation by joseph wiezenbaum. *Technology and Culture*, 17(4):813–816, 1976.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.
- Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*, 2023.
- MinerU Contributors. Mineru: A one-stop, open-source, high-quality data extraction tool. <https://github.com/opendatalab/MinerU>, 2024.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020. doi: 10.1007/s11023-020-09548-1.
- Sarah R I Foot and Nancy Partner. *The sage handbook of historical theory*: Sage publications. 2012.
- François Furet. Quantitative history. *Daedalus*, pp. 151–167, 1971.
- Giselle Gonzalez Garcia and Christian Weilbach. If the sources could talk: Evaluating large language models for research assistance in history. *arXiv preprint arXiv:2310.10808*, 2023.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.
- Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.
- Gabriel et al. Gutherz. Translating akkadian to english with neural machine translation. *PNAS Nexus*, 2023. URL <https://doi.org/10.1093/pnasnexus/pgad096>.
- Lars Peter Hansen and Coauthors. Large language models for economic research: Four key questions. *Centre for Economic Policy Research (CEPR)*, 2023. URL <https://cepr.org/voxeu/columns/large-language-models-economic-research-four-key-questions>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.
- Fei Liu, Dallas Card, and Felix Hamborg. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*, 2023. URL <https://arxiv.org/abs/2403.18932>.
- Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

-
- Benjamin S. Manning, Kehang Zhu, and John J. Horton. Automated social science: Language models as scientists and subjects. *National Bureau of Economic Research (NBER)*, (32381), 2024. doi: 10.3386/w32381. URL <https://arxiv.org/pdf/2404.11794>.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020.
- Qipeng Shi. Rethinking the methodological foundation of historical political science. *Chinese Political Science Review*, 7(1):84–110, 2022.
- Thea Sommerschildt, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747, 2023.
- Joshua Sternfeld. Ai-as-historian. *The American Historical Review*, 128(3):1372–1377, 2023.
- Carla Van Boxtel and Jannet Van Drie. Historical reasoning: Conceptualizations and educational applications. *The Wiley international handbook of history teaching and learning*, pp. 149–176, 2018.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on EMNLP*, pp. 2369–2380, Brussels, Belgium, 2018.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Yifan Zhang, Cheng Wei, Zhengting He, and Wenhao Yu. Geogpt: An assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131:103976, 2024.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

A APPENDIX

A.1 FEATURED SOURCE CODE

A.1.1 BENCHMARK DATA GENERATION

```
1 import json
2 import time
3 import os
4 import re
5 import traceback
6 import difflib
7 import argparse
8 import pdb
9 from utils import openai_chat, qwen_long_chat, deepseek_chat,
    doubao_chat, qwen2_72b_chat
10
11 import random
12 random.seed(37)
13
14
15 def text_similarity(text1, text2):
16     return difflib.SequenceMatcher(None, text1, text2).ratio()
17
18 def replace_with_the_first_source(match):
19     # 获取括号内的内容
20     content = match.group(1)
21     # 查找第一个\史料"m
22     first_material = re.search(r'\史料\d+', content)
23     if first_material:
24         # 返回只包含第一个\史料"的内容m
25         return f'({first_material.group()})'
26     return match.group(0)
27
28
29 def get_full_sources(item_dict, shuffle=False):
30     full_sources_dict = {}
31     full_sources_list = []
32     for item_idx, (_, item) in enumerate(item_dict.items()):
33         for local_idx, source in item['sources'].items():
34             # 合并相同史料的不同表述
35             if source not in full_sources_list:
36                 match_flag = False
37                 for i in range(len(full_sources_list)):
38                     comp_source = full_sources_list[i]
39                     if text_similarity(source, comp_source) >= 0.4: # same as
40                         an existing source
41                         match_flag = True
42                         full_sources_dict[source] = {"global_idx": i,
43                                                     "item_idx": item_idx, "local_idx": local_idx,
44                                                     'redirect': None} # reuse the existing index of
45                         the same source
46
47             # keep the longer one
48             if len(source) > len(comp_source):
49                 full_sources_list[i] = source
50                 full_sources_dict[comp_source]['redirect'] = source
51             else:
52                 full_sources_dict[source]['redirect'] = comp_source
53                 break
54     if not match_flag:
55         full_sources_list.append(source)
56         full_sources_dict[source] = {"global_idx":
57                                     len(full_sources_list), "item_idx": item_idx,
```

```

        "local_idx": local_idx, 'redirect': None} # create a
        new index
53
54 if shuffle:
55     random.shuffle(full_sources_list)
56
57 # align gloabl idx across all same-source sources
58 for source in full_sources_dict.keys():
59     target_source = full_sources_dict[source]['redirect']
60     if target_source is None:
61         full_sources_dict[source]['global_idx'] =
            full_sources_list.index(source)
62     else:
63         # 当同一史料出现三次以上时, 可能存在多次重定向
64         while full_sources_dict[target_source]['redirect'] is not None:
65             target_source = full_sources_dict[target_source]['redirect']
66             full_sources_dict[source]['global_idx'] =
                full_sources_list.index(target_source)
67
68     return full_sources_dict, full_sources_list
69
70 def process_chains(raw_chains, global_sources, strict=True):
71     structured_chains = []
72     deductions = dict()
73     for i, line in enumerate(raw_chains):
74         if len(line)!=2:
75             print('incorrect format')
76             print(line)
77
78         src, dst = line
79         sources = src.replace(' ', '').split('+')
80         if strict:
81             assert (len(sources)<=4 and len(sources)>1), f'推理过程的左侧节点至少
                为个, 最多有个, 现在有24{len(sources)}个}'
82         chain_inputs = []
83         for source in sources:
84             if '(史料' in source:
85                 content, source_idx = re.split(r' (史料[() ', source)
86                 source_idx = int(re.split(r' [)]', source_idx)[0])
87                 assert (source_idx-1 in global_sources.keys()),
                    f"source_idx={source_idx},
                    global_sources={global_sources}"
88                 chain_inputs.append(('史料', source_idx, content))
89             elif '推论' in source:
90                 deduction_idx = int(source.split(f'推论')[1][0])
91                 assert (deduction_idx in deductions.keys()),
                    f"deduction_idx={deduction_idx}, deductions={deductions}"
92                 content = deductions[deduction_idx]
93                 chain_inputs.append(('推论', deduction_idx, content))
94             else:
95                 raise NotImplementedError
96
97         if '推论' in dst:
98             if '(推论' in dst and dst.strip().startswith('推论'):
99                 chain_output = dst.split(' (推论')[0]
100                 chain_output = re.split(r'推论\d', chain_output)[1:]
101             elif '(推论' in dst:
102                 chain_output = dst.split(' (推论')[0]
103             elif dst.strip().startswith('推论'):
104                 chain_output = re.split(r'推论\d', dst)[1:]
105             deduction_idx = int(dst.split('推论')[1][0])
106             assert (deduction_idx<=4)
107             if deduction_idx not in deductions.keys():
108                 deductions[deduction_idx] = chain_output

```

```

109         else:
110             raise NotImplementedError # deductions should not repeat
111     else:
112         assert(i == len(raw_chains)-1)
113         dst = dst.replace('#####', '').replace('\n', '')
114         dst = dst.replace('当前观点', '作者观点') # 检查过程可能出的错误
115         if '(作者观点' in dst:
116             chain_output = dst.split('(作者观点')[0]
117         elif '(反驳错误观点' in dst:
118             chain_output = dst.split('(反驳错误观点')[0]
119         else:
120             raise NotImplementedError
121             # chain_output = dst.split('')[0]
122     structured_chains.append({"input": chain_inputs, 'output':
123                               chain_output})
124     return structured_chains, deductions
125 def all_in_one_true_item_post_process(item):
126     """
127     {
128         "sources": {1: "xxx", 2: "xxx", ...},
129         "deductions": {1: "xxx", 2: "xxx", ...},
130         "option": 作者观点"",
131         "chains": [
132             {"input": 史料推论(['/', idx, content)]},
133         ],
134         "raw_chains": 未结构化的历史推理过程""
135     }
136     """
137     true_item_dict_by_option = {}
138     options = re.split(r'###\s*\观点\d\s*\n###', item)
139     options = [option for option in options if len(option.strip())>0]
140     for option_idx, option in enumerate(options[:4]):
141         item_dict = {'sources': {}, 'deductions': {}, 'chains': []}
142         sections = re.split(r'[\d]', option)[1:]
143         option = sections[0].split('作者观点\n')[1].strip()
144         for existing_option in true_item_dict_by_option.keys():
145             sim = text_similarity(option, existing_option)
146             if sim>=0.7:
147                 print(f"观点过于趋同, 丢弃\nnew:{option}\nold:{existing_option}\nsim={sim}")
148                 continue
149
150         item_dict['option'] = option
151         historical_sources = re.findall(r'史料\d+(?:\n: ||:)(.*?)(?=\史料\n|\Z)', sections[1], re.DOTALL)
152         assert(len(historical_sources)>=2), '史料数量必须>=2'
153         for i, source in enumerate(historical_sources):
154             item_dict['sources'][i] = source.strip()
155
156         # Extract inference chains from section 2 历史推理过程()
157         raw_chains_txt = sections[2].split('历史推理过程\n')[1].replace('(', ' (').replace(')', ')')
158         raw_chains_txt = re.sub(r'() ([^+])',
159                                replace_with_the_first_source, raw_chains_txt) # 多史料节点中只保留一个节点
159         item_dict['raw_chains'] = raw_chains_txt.replace('\n\n',
160                                                         '\n').replace('#####', '').split('(作者观点)')[0] + '(作者观点)'
161         raw_chains = [line.split('->') for line in
162                       raw_chains_txt.split('\n') if '->' in line]
163         structured_chains, deductions = process_chains(raw_chains,
164                                                         item_dict['sources'], strict=False)
165         item_dict['chains'] = structured_chains

```

```

163     item_dict['deductions'] = deductions
164     assert(' (作者观点)' in item_dict['raw_chains'])
165     item_dict['item_idx'] = option_idx
166     item_dict['label'] = True
167     true_item_dict_by_option[item_dict['option']] = item_dict
168
169     return true_item_dict_by_option
170
171 def true_item_post_process(item):
172     """
173     {
174         "sources": {1: "xxx", 2: "xxx", ...},
175         "deductions": {1: "xxx", 2: "xxx", ...},
176         "option": 作者观点",
177         "chains": [
178             {"input": 史料推论(['/', idx, content])}
179         ],
180         "raw_chains": 未结构化的历史推理过程"
181     }
182     """
183     item_dict = {'sources': {}, 'deductions': {}, 'chains': []}
184     sections = re.split(r'[\d]', item)[1:]
185     item_dict['option'] = sections[0].split('作者观点\n')[1].strip()
186
187     historical_sources = re.findall(r'史料\d+(?:\n:|:)(.*?) (?:\s史
188         料\n|\Z)', sections[1], re.DOTALL)
189     assert(len(historical_sources)>=2), '史料数量必须>=2'
190     for i, source in enumerate(historical_sources):
191         item_dict['sources'][i] = source.strip()
192
193     # Extract inference chains from section 2 历史推理过程()
194     raw_chains_txt = sections[2].split('历史推理过
195         程\n')[1].replace('(', '(').replace(')', ')')
196     raw_chains_txt = re.sub(r'() ([^+])',
197         replace_with_the_first_source, raw_chains_txt) # 多史料节点中只保留一
198         个节点
199     item_dict['raw_chains'] = raw_chains_txt.replace('\n\n',
200         '\n').replace('#####', '').split(' (作者观
201         点)')[0] + ' (作者观点)'
202     raw_chains = [line.split('->') for line in raw_chains_txt.split('\n')
203         if '->' in line]
204     structured_chains, deductions = process_chains(raw_chains,
205         item_dict['sources'], strict=False) # true 较昂贵, 尽可能只跑一次, 出错
206         了由环节修复。optioncheck
207     item_dict['chains'] = structured_chains
208     item_dict['deductions'] = deductions
209     assert(' (作者观点)' in item_dict['raw_chains'])
210
211     return item_dict
212
213 def false_item_post_process(false_item, source_dict):
214     item_dict = {'sources': None, 'deductions': {}, 'chains': []}
215
216     sections = re.split(r'[\d]', false_item)[1:]
217     item_dict['option'] = sections[0].split('错误观点\n')[1].strip()
218
219     # Extract inference chains from section 2 证据链推理过程()
220     raw_chains_txt = sections[1].split('历史推理过
221         程\n')[1].replace('(', '(').replace(')', ')')
222     raw_chains_txt = re.sub(r'() ([^+])',
223         replace_with_the_first_source, raw_chains_txt) # 多史料节点中只保留一
224         个节点

```

```

214 item_dict['raw_chains'] = raw_chains_txt.replace('\n\n',
215         '\n').replace('#####', '').split('(反驳错误观
216 点)')[0] + '(反驳错误观点)'
217 raw_chains = [line.split('->') for line in raw_chains_txt.split('\n')]
218 if '->' in line]
219 structured_chains, deductions = process_chains(raw_chains,
220         source_dict)
221 item_dict['chains'] = structured_chains
222 item_dict['deductions'] = deductions
223 assert('(反驳错误观点)' in item_dict['raw_chains'])
224
225 return item_dict
226
227 def item_checking_post_process(item_dict, check_res, source_dict):
228     sections = re.split(r'[\d]', check_res)[2:]
229     if '无需修改' not in sections[0]:
230         label = '必须仍然是正确观点' if item_dict['label'] else '必须仍然是有辨别
231         难度的错误观点'
232         new_option = sections[0].split(f'修改后的观点 ({label})')[1].strip()
233         assert('label' not in new_option and '\n' not in
234             new_option.strip()), f'可能是非法的修
235             改: {new_option}'
236         item_dict['option'] = new_option
237     if '无需修改' not in sections[1]:
238         raw_chains_txt = sections[1].split('修改后的历史推理过
239         程\n')[1].replace('(', '(').replace(')', ')')
240         raw_chains_txt = re.sub(r' () ([^ ]+)',
241             replace_with_the_first_source, raw_chains_txt) # 多史料节点中只保
242             留一个节点
243         item_dict['raw_chains'] = raw_chains_txt.replace('(',
244             ' ').replace(')', ' ').replace('\n\n',
245             '\n').replace('#####', '')
246         raw_chains = [line.split('->') for line in
247             raw_chains_txt.split('\n')] if '->' in line]
248         structured_chains, deductions = process_chains(raw_chains,
249             source_dict)
250         item_dict['chains'] = structured_chains
251         item_dict['deductions'] = deductions
252
253     if item_dict['label'] == True:
254         assert('(作者观点)' in item_dict['raw_chains'])
255         item_dict['raw_chains'] = item_dict['raw_chains'].split('(作者观
256         点)')[0] + '(作者观点)'
257     else:
258         assert('(反驳错误观点)' in item_dict['raw_chains'])
259         item_dict['raw_chains'] = item_dict['raw_chains'].split('(反驳错误观
260         点)')[0] + '(反驳错误观点)'
261
262     return item_dict
263
264 def step_by_step_generate(md_file, chat_api='openai', verbose=False):
265     if chat_api == 'openai':
266         chat = openai_chat
267         true_option_max_retries = 3
268         false_option_max_retries = 3
269         check_option_max_retries = 5
270     elif chat_api == 'qwen':
271         chat = qwen_long_chat
272         true_option_max_retries = 3
273         false_option_max_retries = 3
274         check_option_max_retries = 5
275     elif chat_api == 'deepseek':

```

```

263     chat = deepseek_chat
264     true_option_max_retries = 3
265     false_option_max_retries = 3
266     check_option_max_retries = 5
267 elif chat_api == 'doubao':
268     chat = doubao_chat
269     true_option_max_retries = 3
270     false_option_max_retries = 3
271     check_option_max_retries = 5
272 elif chat_api == 'qwen2':
273     chat = qwen2_72b_chat
274     true_option_max_retries = 3
275     false_option_max_retries = 3
276     check_option_max_retries = 5
277 else:
278     raise NotImplementedError
279
280 paper_md = open(md_file, 'r').read()
281
282 cur_options_txt = '暂无'
283 option_mapper = {0: 'A', 1: 'B', 2: 'C', 3: 'D', 4: 'E', 5: 'F', 6:
284                 'G', 7: 'H'}
285 true_item_dict_by_option = {}
286 false_item_dict_by_option = {}
287 merged_item_dict_by_option = {}
288 output_item_dict_by_option = {}
289
290 # TODO, currently we only sample one combination (totally 70
291     combinations)
292 COMBINATIONAL_GENERATION_FOR_TRUE_OPTIONS = False
293
294 ##### 1. 抽取正确选项
295 # 总是先生成个正确选项，保留这些正确选项的所有史料，但每一次只使用部分正确选项来构成
296     题目4
297 if verbose:
298     print('>>> Step 01. 生成正确选项...')
299
300 num_options = 4
301 true_reasoning_prompt =
302     open('prompts/step_by_step_prompt_02_true_reasoning_extraction.txt',
303         'r').read()
304 all_in_one_true_reasoning_prompt =
305     open('prompts/step_by_step_prompt_' \
306         '02_all_in_one_true_reasoning_extraction.txt', 'r').read()
307 ALL_IN_ONE_TRUE_OPTION_GEN = True
308 if ALL_IN_ONE_TRUE_OPTION_GEN:
309     messages = [
310         {"role": "system", "content":
311             all_in_one_true_reasoning_prompt},
312         {"role": "user", "content": f'论文内
313             容[1]\n\n'"{paper_md}"'\n\n'}
314     ]
315     count = 0
316     true_item_dict_by_option = None
317     while count < true_option_max_retries:
318         try:
319             all_in_one_true_items = chat(messages, max_tokens=3000,
320                 verbose=verbose)
321             true_item_dict_by_option =
322                 all_in_one_true_item_post_process(all_in_one_true_items)
323             break
324         except Exception as e:
325             print(f"第{count}次抽取正确选项时捕获到异常+1: {e}")
326             # traceback.print_exc()
327             count += 1

```



```

318     assert(true_item_dict_by_option is not None), '无法抽取正确选项'
319 else:
320     for item_idx in range(num_options):
321         messages = [
322             {"role": "system", "content": true_reasoning_prompt},
323             {"role": "user", "content": f'论文内
324                 容[1]\n\n""{paper_md}""\n\n现有作者观
325                 点[2]\n{cur_options_txt}'}
326         ]
327         count = 0
328         true_item_dict = None
329         while count < true_option_max_retries:
330             try:
331                 true_item = chat(messages, max_tokens=1200,
332                                 verbose=verbose)
333                 true_item_dict = true_item_post_process(true_item)
334                 for existing_option in true_item_dict_by_option.keys():
335                     sim = text_similarity(true_item_dict['option'],
336                                         existing_option)
337                     assert(sim <= 0.7), f"观点过于趋同, 重新生成" \
338                                     "\nnew:{true_item_dict['option']}" \
339                                     "\nold:{existing_option}\nsim={sim}"
340                     break
341             except Exception as e:
342                 print(f"第{count}次抽取正确选项时捕获到异常+1: {e}")
343                 # traceback.print_exc()
344                 count += 1
345             if true_item_dict is None:
346                 print('difficult to extract more options')
347                 continue
348
349             true_item_dict['item_idx'] = item_idx
350             true_item_dict['label'] = True
351             option = true_item_dict['option']
352             if cur_options_txt == '暂无':
353                 cur_options_txt = f'{option_mapper[item_idx]}. {option}\n'
354             else:
355                 cur_options_txt += f'{option_mapper[item_idx]}. {option}\n'
356             true_item_dict_by_option[option] = true_item_dict
357             assert(len(true_item_dict_by_option) > 0), '无法抽取正确选项'
358
359 # 整合正确选项信息的所有史料, 但仅使用部分正确选项
360 _, full_sources_list = get_full_sources(true_item_dict_by_option,
361                                         shuffle=True)
362 new_sources_txt = ''
363 for i, source in enumerate(full_sources_list):
364     new_sources_txt += f'史料{i: +1}\n{source}\n'
365 cur_options_txt = '' # reset cur_options
366 for option_idx, option in enumerate(true_item_dict_by_option.keys()):
367     cur_options_txt += f'{option_mapper[option_idx]}. {option}\n'
368 tmp_source_dict = {i: source for i, source in
369                   enumerate(full_sources_list)}
370
371 ##### 2. 生成错误选项
372 if verbose:
373     print('>>> Step 02. 生成错误选项...')
374
375 false_reasoning_prompt =
376     open('prompts/step_by_step_prompt_03_false_reasoning_generation.txt',
377         'r').read()
378 for _ in range(num_options-1): # at most 3 false options
379     messages = [
380         {"role": "system", "content": false_reasoning_prompt},

```

```

374         {"role": "user", "content": f"#####\nInput:\n史
          料列表[1]\n\n{new_sources_txt}\n现有观
          点[2]\n{cur_options_txt}"
375     ]
376
377     count = 0
378     false_item_dict = None
379     while count < false_option_max_retries:
380         try:
381             false_item = chat(messages, max_tokens=1200, verbose=verbose)
382             false_item_dict = false_item_post_process(false_item,
                 tmp_source_dict)
383             for existing_option in false_item_dict_by_option.keys():
384                 sim = text_similarity(false_item_dict['option'],
                     existing_option)
385                 assert(sim <= 0.7), f"观点过于趋同, 重新生成" \
386                     "\nnew:{false_item_dict['option']}" \
387                     "\nold:{existing_option}\nsim={sim}"
388                 break
389             except Exception as e:
390                 print(f"第{count}次生成错误选项时捕获到异常+1: {e}")
391                 # traceback.print_exc()
392                 count += 1
393             if false_item_dict is None:
394                 raise NotImplementedError('无法生成错误选项...')
395
396             false_item_dict['item_idx'] = len(true_item_dict_by_option) +
                 len(false_item_dict_by_option)
397             false_item_dict['label'] = False
398             false_item_dict['sources'] = {i: source for i, source in
                 enumerate(full_sources_list)}
399             option = false_item_dict['option']
400             cur_options_txt += f'{option_mapper[len(true_item_dict_by_option)
                 + len(false_item_dict_by_option)]}. {option}\n'
401             false_item_dict_by_option[option] = false_item_dict
402
403
404             for option in true_item_dict_by_option.keys():
405                 merged_item_dict_by_option[option] =
                     true_item_dict_by_option[option]
406             for option in false_item_dict_by_option.keys():
407                 merged_item_dict_by_option[option] =
                     false_item_dict_by_option[option]
408
409             ##### 3. 检查
410             if verbose:
411                 print('>>> Step 03. 检查选项...')
412                 print('merged_item_dict_by_option', merged_item_dict_by_option)
413                 # check_reasoning_prompt =
414                 open('prompts/step_by_step_prompt_04_reasoning_checking.txt',
                     'r').read()
415                 true_check_reasoning_prompt =
416                 open('prompts/step_by_step_prompt_04_true_reasoning_checking.txt',
                     'r').read()
417                 false_check_reasoning_prompt =
418                 open('prompts/step_by_step_prompt_04_false_reasoning_checking.txt',
                     'r').read()
419                 final_full_sources_dict, final_full_sources_list =
420                 get_full_sources(merged_item_dict_by_option, shuffle=True)
421                 final_sources_txt = ''
422                 for i, key in enumerate(final_full_sources_list):
423                     final_sources_txt += f'史料{i: +1}\n{key}\n\n'
424
425             # deprecated: check for several times
426             for check_i in range(1):

```

```

423 new_merged_item_dict_by_option = {}
424 for option_idx, (option, item) in
    enumerate(merged_item_dict_by_option.items()):
425
426     print(f"检查次数: {check_i, 选
        项: +1}{option_mapper[option_idx, ]: label[item['label']]
        ...}")
427     raw_chains = item['raw_chains']
428     for i, source in item['sources'].items(): # since 0
429         local_name = f' (史料{i} +1)'
430         global_idx = final_full_sources_dict[source]['global_idx']
431         global_name = f' (史料{global_idx} +1)'
432         raw_chains = raw_chains.replace(local_name, global_name)
433
434     if item['label']:
435         check_reasoning_prompt = true_check_reasoning_prompt
436         this_input = f"史料列表[1]\n{final_sources_txt.strip()}\n\n当前
            观点 (正确观点) [2]\n{option.strip()}\n\n当前观点的历史推理过
            程[3]\n{raw_chains.strip()}\n\n"
437     else:
438         check_reasoning_prompt = false_check_reasoning_prompt
439         this_input = f"史料列表[1]\n{final_sources_txt.strip()}\n\n当前
            观点 (有辨别难度的错误观点) [2]\n{option.strip()}\n\n反驳当前观
            点的历史推理过程[3]\n{raw_chains.strip()}\n\n"
440
441     messages = [
442         {"role": "system", "content": check_reasoning_prompt},
443         {"role": "user", "content":
            f"#####\nInput:\n{this_input}"}
444     ]
445
446     count = 0
447     output_item_dict = None
448     while count < check_option_max_retries:
449         try:
450             check_res = chat(messages, max_tokens=1200,
                verbose=verbose)
451             output_item_dict = item_checking_post_process(item,
                check_res, tmp_source_dict)
452             break
453         except Exception as e:
454             print(f"第{count}次检查选项时捕获到异常+1): {e}")
455             # traceback.print_exc()
456             count += 1
457     if output_item_dict is None:
458         raise NotImplementedError('无法检查选项...')
459     if verbose:
460         print('this_input', this_input)
461         print('check_res', check_res)
462     new_merged_item_dict_by_option[output_item_dict['option']] =
        output_item_dict
463     merged_item_dict_by_option = new_merged_item_dict_by_option
464
465     if verbose:
466         print('>>> Step 04. 生成题目...')
467         print('new_merged_item_dict_by_option', merged_item_dict_by_option)
468
469     num_true_items = random.randint(1, min(len(true_item_dict_by_option),
        num_options)) # 可能存在不足个的情
        况4
470     num_false_items = num_options - num_true_items
471     new_true_item_dict_by_option = {}
472     new_false_item_dict_by_option = {}
473     for option, option_v in merged_item_dict_by_option.items():
474         if option_v['label'] == True:

```

```

475         new_true_item_dict_by_option[option] = option_v
476     else:
477         new_false_item_dict_by_option[option] = option_v
478
479     sampled_true_options =
480         random.sample(list(new_true_item_dict_by_option.keys()),
481                        num_true_items)
482     sampled_false_options =
483         random.sample(list(new_false_item_dict_by_option.keys()),
484                        num_false_items)
485
486     for option in sampled_true_options + sampled_false_options:
487         output_item_dict_by_option[option] =
488             merged_item_dict_by_option[option]
489         output_item_dict_by_option[option] =
490             merged_item_dict_by_option[option]
491
492     # 6. 生成题目
493     final_sources_txt = final_sources_txt
494     final_options = list(output_item_dict_by_option.keys())
495     random.shuffle(final_options)
496     final_options_txt = ''
497     gt_answer = ''
498     final_chains = ''
499     for option_idx, option in enumerate(final_options):
500         final_options_txt += f'{option_mapper[option_idx]}. {option}\n'
501         this_final_chains =
502             output_item_dict_by_option[option]['raw_chains'].strip()
503         idx = option_mapper[option_idx]
504         if output_item_dict_by_option[option]['label']:
505             gt_answer += idx
506             assert('作者观点' in this_final_chains)
507             this_final_chains = this_final_chains.replace('作者观点', f'选
508                 项{idx}正确')
509         else:
510             assert('反驳错误观点' in this_final_chains)
511             this_final_chains = this_final_chains.replace('反驳错误观点', f'选
512                 项{idx}错误')
513
514         final_chains += f'{option_idx+1}. 选
515             项{idx: } \n{this_final_chains} \n \n'
516
517     output_txt = f"""不定项选择题阅读下列史料，判断下列选项哪些最有可能成立：
518
519     {final_sources_txt}选项：
520
521     {final_options_txt}这是一道不定项选择题，至少有一个正确选项，至多有四个正确选项。仅
522     回复所有正确选项的字母编号，不要回复其他内容：
523
524     #####正确答案：
525
526     {gt_answer}历史推理过程：
527
528     {final_chains}
529     """
530     if verbose:
531         print(f'\n\nOutput Text:\n{output_txt}')
532     output_dict = {}
533     output_dict['source_path'] = md_file
534     output_dict['chat_api'] = chat_api

```

```

529 output_dict['output_txt'] = output_txt
530 output_dict['final_options'] = final_options
531 output_dict['output_item_dict_by_option'] = output_item_dict_by_option
532 output_dict['true_item_dict_by_option'] = new_true_item_dict_by_option
533 output_dict['false_item_dict_by_option'] =
    new_false_item_dict_by_option
534 output_dict['old_true_item_dict_by_option'] = true_item_dict_by_option
535 output_dict['old_false_item_dict_by_option'] =
    false_item_dict_by_option
536 output_dict['final_full_sources_dict'] = final_full_sources_dict
537
538 return output_dict
539
540
541 def build_args():
542     parser = argparse.ArgumentParser("hist reasoning")
543     parser.add_argument("-s", "--start-index", default=0, type=int,
544                         help='起始样本的序
545                             号')
544     parser.add_argument("-e", "--end-index", default=-1, type=int,
545                         help='结束样本的序
546                             号')
545     parser.add_argument("--split", default=2023, type=int, choices=[2021,
546                             2022, 2023, 2024], help='数据集编
547                             号')
546     parser.add_argument("--rewrite", action="store_true")
547
548     args = parser.parse_args()
549     return args
550
551
552 if __name__ == '__main__':
553     args = build_args()
554     print(args)
555     split = f'md_{args.split}'
556     data_path = f'./data/inputs/{split}'
557     output_path = f'./data/outputs/{split}'
558     verbose = True
559     chat_apis = ['openai']
560     # chat_apis = ['deepseek']
561     # chat_apis = ['doubao']
562     # chat_apis = ['qwen2']
563
564     if os.path.exists(output_path) is False:
565         os.makedirs(output_path)
566     for chat_api in chat_apis:
567         this_output_path = os.path.join(output_path, chat_api)
568         if os.path.exists(this_output_path) is False:
569             os.makedirs(this_output_path)
570
571     files = os.listdir(data_path)
572     files = [file for file in files if file.endswith('.md')]
573     raw_len = len(files)
574
575     assert(args.start_index>=0)
576     if args.end_index == -1:
577         files = files[args.start_index:]
578     else:
579         assert(args.end_index > args.start_index)
580         files = files[args.start_index:args.end_index]
581
582     for file_idx, file in enumerate(files):
583         for chat_api in chat_apis:
584             read_idx = args.start_index+file_idx+1
585             this_output_path = os.path.join(output_path, chat_api)

```

```

586 data_file = os.path.join(data_path, file)
587 output_file = f'{read_idx}-{file}'
588 output_json_path = os.path.join(this_output_path,
589 output_file).replace('.md', '.json')
590 output_txt_path = os.path.join(this_output_path,
591 output_file).replace('.md', '.txt')
592
593 print(f'Generating {output_txt_path}
594       ({args.start_index+file_idx+1}/{raw_len})\n\n' \
595       '#####\n\n')
596 start_time = time.time()
597 if args.rewrite == False and os.path.exists(output_txt_path) ==
598 True:
599     print(f'Rewrite=False and {output_txt_path} already existed.
600           Continue...')
601     continue
602 else:
603     try:
604         output_dict = step_by_step_generate(data_file, chat_api,
605                                             verbose=verbose)
606     except Exception as e:
607         print(f"[Error当处理]{output_txt_path}时遇到报错, 跳过: } {e}")
608         # traceback.print_exc()
609         continue
610
611     with open(output_json_path, 'w') as w:
612         w.write(json.dumps(output_dict, indent=4,
613                           ensure_ascii=False))
614     with open(output_txt_path, 'w') as w:
615         w.write(output_dict['output_txt'])
616
617     timecost = time.time() - start_time
618     print(f'Progress: {read_idx}/{raw_len}, api: {chat_api},
619           timecost: {timecost:.2f}s' \
620           ', written to {output_txt_path}\n\n' \
621           '#####\n\n')

```

A.1.2 HR-TUNING DATA GENERATION

```

1 import json
2 import time
3 import os
4 import re
5 import random
6 import traceback
7 import difflib
8 import argparse
9 import pdb
10
11 def build_args():
12     parser = argparse.ArgumentParser("hist reasoning")
13     parser.add_argument("--train-splits", default='2021-2023', type=str,
14                         help='数据集')
15     parser.add_argument("--valid-percentage", default=0., type=float,
16                         help='数据集')
17     parser.add_argument("--method", default='options',
18                         choices=['options', 'source_understanding', 'source_explanation',
19                                 'source_deduction'], type=str, help='任务组合')
20     parser.add_argument("--source-shuffle", action="store_true")

```

```

17 parser.add_argument("--no-source-idx", action="store_true")
18 parser.add_argument("--prompts", default='1', choices=['1', '2', '3',
    '1+2', '1+3', '2+3', '1+2+3'], type=str, help='提示词组合, 对应基本提
    示词, 对应带有简单背景的提示词, 对应带有细致背景的提示词123')
19
20 args = parser.parse_args()
21 return args
22
23
24 def generate_sft_data(args):
25     prompt_ids = args.prompts.split('+')
26     prompts = []
27     for prompt_id in prompt_ids:
28         prompt_id = int(prompt_id)
29         prompt_file = os.path.join(
30             f'prompts/sft_prompts/sft_prompt_{prompt_id:02d}_mod.txt')
31         prompts += [open(prompt_file).read()]
32
33     start_split, end_split = args.train_splits.split('-')
34     start_split = int(start_split)
35     end_split = int(end_split)
36     assert(end_split >= start_split and start_split in [2021, 2022, 2023,
    2024] and end_split in [2021, 2022, 2023, 2024])
37
38     output_data = []
39     explanation_output_data = []
40     deduction_output_data = []
41     cases = dict()
42     insert_option = args.method == 'options'
43     use_source = 'source_' in args.method
44     if args.method == 'source_understanding':
45         use_source_explanation = True
46         use_source_deduction = True
47     if args.method == 'source_explanation':
48         use_source_explanation = True
49         use_source_deduction = False
50     if args.method == 'source_deduction':
51         use_source_explanation = False
52         use_source_deduction = True
53     for split in range(start_split, end_split+1):
54         data_path = f'./data/outputs/md_{split}/openai'
55
56         files = os.listdir(data_path)
57         files = sorted([file for file in files if file.endswith('.json')])
58         for file in files:
59             with open(os.path.join(data_path, file), 'r') as r:
60                 data = json.loads(r.read())
61                 case = data['output_txt']
62                 sources = data['output_txt'].split('阅读下列史料, 判断下列选项哪些最有
    可能成立: \n\n')[1].split('选项: ')[0].strip()
63
64                 if use_source_explanation or use_source_deduction:
65                     explanation_prompt = open(
66                         './prompts/sft_prompts/sft_prompt_source_explanation.txt',
67                         'r').read()
68                     deduction_prompt = open(
69                         './prompts/sft_prompts/sft_prompt_source_deduction.txt',
70                         'r').read()
71                     answer = case.split('#####')[1]
72                     chains = answer.split('历史推理过程:')[1].strip().split('\n')
73                     for row in chains:
74                         source_splits = re.split(r'史料\d: +', sources)
75                         source_splits = [split.strip() for split in source_splits]
76                         source_splits = [split for split in source_splits if
77                             len(split)>0]

```



```

75     if '->' in row and '选项' not in row and '推
76         论' not in row.split('->')[0]:
77         explanations = row.replace('
78             ', '').split('->')[0].strip().split('+')
79         deduction_output = row.replace('
80             ', '').split('->')[1].split(' (推
81             论') [0].strip()
82         comb_sources = []
83         new_chain_txts = []
84         explanation2source = {}
85         source2idx = {}
86         try:
87             for i, explanation in enumerate(explanations):
88                 content, source_idx = re.split(r' (史
89                     料[()]+, explanation
90                     source_idx = int(re.split(r' [()]+',
91                         source_idx) [0])
92                 this_source = source_splits[source_idx-1]
93                 if use_source_explanation:
94                     explanation_output_data.append({"instruction":
95                         explanation_prompt, "input": this_source,
96                         "output": content})
97                     explanation2source[content] = this_source
98                     comb_sources.append(this_source)
99
100                comb_sources_txt = ''
101                if args.source_shuffle:
102                    random.shuffle(comb_sources)
103                for i, txt in enumerate(comb_sources):
104                    source2idx[txt] = i
105                    comb_sources_txt += f'史料{i: +1}\n{txt}\n\n'
106
107                for explanation in explanation2source.keys():
108                    source_idx =
109                        source2idx[explanation2source[explanation]]
110                    # shuffled_idx
111                    if not args.no_source_idx:
112                        new_chain_txts.append(f'{explanation (史
113                            料){source_idx} +1}')
114                    else:
115                        new_chain_txts.append(f'{explanation}')
116
117                new_chain_txt = ' + '.join(new_chain_txts)
118                new_chain_txt += f' -> {deduction_output (推论) }'
119
120                if use_source_deduction:
121                    deduction_output_data.append({"instruction":
122                        deduction_prompt, "input": comb_sources_txt,
123                        "output": new_chain_txt})
124            except Exception as e:
125                print(f'[Error for row: {row}, {e}']
126            # output_data = explanation_output_data[:5000] +
127                deduction_output_data[:10000]
128            output_data = explanation_output_data + deduction_output_data
129
130            if insert_option:
131                true_item_dict_by_option = data['true_item_dict_by_option']
132                false_item_dict_by_option = data['false_item_dict_by_option']
133
134            cases[sources] = []
135            for option, option_data in true_item_dict_by_option.items():
136                label = '正确'
137                chains = option_data['raw_chains'].replace('作者观点', '观点
138                    正确').strip()
139                cases[sources].append((option, label, chains))

```

```

125         for option, option_data in false_item_dict_by_option.items():
126             label = '错误'
127             chains = option_data['raw_chains'].replace('反驳错误观
128                 点', '观点错误').strip()
129             cases[sources].append((option, label, chains))
130
131     if insert_option:
132         for source, options in cases.items():
133             for (option, label, chains) in options:
134                 input = f"史料列表[1]\n{source}\n当前观点[2]\n{option}\n\n"
135                 output = f"分析过程: [1]\n{chains}\n\n观点判断: [2]\n{label}"
136                 for prompt in prompts:
137                     instruction = prompt
138                     output_data.append({"instruction": instruction, "input":
139                         input, "output": output})
140
141     output_path = f'./data/'
142     if os.path.exists(output_path) is False:
143         os.makedirs(output_path)
144     if use_source_explanation or use_source_deduction:
145         shuffle_suffix = '_shuffle' if args.source_shuffle else ""
146         no_source_idx_suffix = '_no_source_idx' if args.no_source_idx else ""
147
148     output_file = os.path.join(output_path,
149         f"sft_hist_reasoning_{args.train_splits.replace('-', '_')} \
150         f"_{args.method}{shuffle_suffix}{no_source_idx_suffix}.json")
151     else:
152         output_file = os.path.join(output_path,
153             f"sft_hist_reasoning_{args.train_splits.replace('-', '_')} \
154             f"_prompts_{args.prompts.replace('+', '_')}_{args.method}.json")
155     with open(output_file, 'w') as w:
156         w.write(json.dumps(output_data, indent=4, ensure_ascii=False))
157
158     print(f'Finished.\nNum of cases: {len(output_data)}, written to
159         {output_file} ...')
160
161 if __name__ == '__main__':
162     args = build_args()
163     print(args)
164     generate_sft_data(args)

```

A.2 FEATURED PROMPTS

A.2.1 BENCHMARK DATA GENERATION

Generating True Option Candidates

-Goal-

历史是一门注重逻辑推理和严密论证的实证性的人文社会学科。遵循史料研究注意事项，从上传的历史学科论文中，针对作者观点，抽取出相关的史料列表，以及作者基于史料的形式化的历史推理过程。

-Knowledge-

史料研究注意事项：

1. 史料价值分类。第一手史料（直接史料）：原始资料，即历史人物、事件、现象发生的那个时代留下来的材料，特别是当事人和目击者提供的史料，直接反映了所研究对象的情况。第二手材料（间接史料）：距离那个时代较远的记录、转述，或者后人的研究成果，间接反映了研究对象的情况。
2. 对史料要合理利用，做到史论结合论由史出，有一分材料说一分话。
3. 搜集更多的史料，对多种史料进行甄选、辨别，多源互证。

4. 孤证不立：尽可能多地搜集史料。

-Constraints-

1. 你将作者的推理过程转化为证据链，而不是根据史料自行生成证据链。
2. 基于基本逻辑和一定的历史背景知识，人类可以从历史推理过程中推导出作者观点。
3. 史料不允许是作者的观点，必须是一手资料的内容或者之前学者的观点，不要使用作者转述。
4. 史料列表中的每一个史料都必须在证据链推理过程中被使用到，且仅使用一次。
5. 优先使用一手史料，当一手史料用完之后，再考虑使用二手史料。
6. 你生成的作者观点，不允许与给输入的现有观点重复或冲突，不允许在历史推理过程中使用现有观点中的信息。
7. 每个观点的生成是互相独立的，各自须包含作者观点、史料列表和历史推理过程，但您应该使得4个选项都正确但各不相同，它们可以反映作者在不同事件、不同角度、不同分论点上的观点。
8. 不允许无中生有，不允许编造论文中不存在的观点和史料!!!
9. 每次回复中，您须严格遵循下述输出格式，不要回复无关内容。YOU MUST FOLLOW THE OUTPUT FORMAT BELOW.

-Input Format-

论文内容

""""论文内容""""

-Output Format-

您一共要生成**4个观点**，每个选项遵循如下格式：

####

观点n

####

作者观点

从论文中抽取的作者观点，该观点需要是作者使用多个史料，基于逻辑推理和严密论证后得出的观点，允许对该观点进行同义改写，但不允许篡改其本身含义。

史料列表

要求：

- 每个观点利用的史料最少为2个，最多不超过4个；
- 不允许使用上传论文作者的观点/表述作为史料（而只能作为观点），尽可能使用论文中的引用作为史料，尤其是文言文。
- 遵循以下格式：

史料1：

>来自原文的史料内容（不得篡改），尽可能是文言文，一手史料不允许是作者的观点，必须是一手资料的内容，不要使用作者转述；仅包含史料内容，不要加入作者解释

——史料出处，《文献名》或XXX《文献名》

史料2：

>来自原文的史料内容（不得篡改），尽可能是文言文，二手史料不允许是作者的观点，必须是之前学者的观点，仅包含史料内容，不要加入作者解释

——史料出处，《文献名》或XXX《文献名》

历史推理过程

从上传论文中抽取作者观点（允许改写但不得歪曲其含义），以及作者推导出该观点使用的历史推理过程，该历史推理过程至少需要使用到2个史料。注意，历史推理过程只能使用史料和基于史料的推论作为输入，请勿使用现有观点！历史推理过程包含单步推理或多步推理两种可能情况：

1. 单步推理：每一行仅允许有一个推理符“->”，该推理符左侧**至少有2个节点**，最多不超过4个，且均为史料；该推理符右侧**仅有1个节点**，且必然为作者观点；
史料1的解释（史料1）+ 史料2的解释（史料2）+ ... ->作者观点的内容（作者观点）

2. 多步推理：每一行仅允许有一个推理符“->”，每一个推理符左侧的节点**至少有2个节点**，最多不超过4个，这些节点可能是史料，也可能是已有的推论；推理符右侧**仅有1个节点**，非最后一行的右侧节点

必然是新的推论，最后一个右侧节点必然为作者观点；
史料1的解释（史料1）+ 史料2的解释（史料2）->推论1的内容（推论1）
史料3的解释（史料3）+ 史料4的解释（史料4）->推论2的内容（推论2）
推论1+ 推论2->作者观点的内容（作者观点）

推理约束：

- 您应当确保，您每一行历史推理过程都要符合历史常识和基本逻辑，您的历史推理过程仅依赖于史料、历史常识和基本逻辑，后续用户是无法看到您的论文内容的。
- “->”左侧的每个史料，都应该对推导出“->”右侧的推论有重要作用。
- 推论从1开始计数，新的推论在原有最大推论序号上+1，推论的序号不允许重复和跳过。
- 禁止使用现有作者观点进行历史推理！！
- 基于史料生成的推论，至少需要被作为输入使用一次，不允许仅生成而不被使用。
- 禁止根据单个史料、或者单个推论，来推理出新的推论或作者观点。
- 禁止一次性解释多个史料，如：
 - 这是错误的！“史料m和史料n的解释（史料m和史料n）->新的节点“，这是错误的！
 - 这是错误的！“史料m和史料n的解释（史料m与史料n）->新的节点“，这是错误的！
 - 这是错误的！“史料m和史料n的解释（史料m，史料n）->新的节点“，这是错误的！
 - 这是错误的！“史料m和史料n的解释（史料m+史料n）->新的节点“，这是错误的！
- 每一行仅有一个推理符“->“！
- 尽可能保持史料解释和推论内容的简洁。

#####

-Examples-

#####

Input:

论文内容

“””论文内容“””

#####

Output:

####

观点1

####

作者观点

晚清山西土地清丈活动的成功在于地方绅士与地方官员的合作，以及从解州到阳曲的经验传播和技术人员的流动形成了清丈官员网络圈。

史料列表

史料1:

>清丈活动中，全县235个村庄被分为12都，每都编造鱼鳞册10册，根据粮地的多少，或一村为一册，或一村为两册，粮地少者数村为一册。“分清围数，编明段落，其四至、粮额，俱各注明，一地一串，编定号数，不得混淆”，最终形成的鱼鳞册以四至明确的田地为主要内容。

——《西段屯地理图》

史料2:

>光绪五年的解州清丈办法创之于当地乡绅孙守恒，推广至山西阳曲、陕西朝邑等地，未受介休清丈办法的直接影响。

——《高崇基行之于介休而效，马丕瑶行之于解州而益效》

史料3:

>巡抚张之洞面对各地不断的荒地请豁，提议通过土地清丈解决田赋积弊。但此次清丈并非制度化地在全省推行，而采取自愿原则，“此后晋省州县，何处能办清丈者，由省委员勘明举办”。

——《张之洞全集》

史料4:

>以解州绅士孙守恒为代表、掌握清丈和编修鱼鳞册专业技术的地方绅士群体，通过有组织、跨地域活动，为各地清丈实践提供专业技术和人员保障。

——山西大学中国社会史研究中心教授胡英泽

历史推理过程

全县235个村庄编造鱼鳞册，形成明确的田地内容（史料1）+解州清丈推广至阳曲、陕西朝邑等地（史料2）

->地方绅士在土地清丈中起到了关键作用，并与地方官员合作形成有效的清丈方法（推论1）

张之洞提倡自愿清丈，并在多地开展清丈（史料3）+地方绅士通过跨地域活动促进清丈实践（史料4）->成功的清丈活动离不开地方官员与绅士的合作，并形成了清丈官员网络圈，从解州推广到阳曲等地（推论2）

推论1+推论2->晚清山西土地清丈活动的成功在于地方绅士与地方官员的合作，以及从解州到阳曲的经验传播和技术人员的流动形成了清丈官员网络圈（作者观点）

####

观点2

####

作者观点

解州、阳曲等地的土地清丈在应对自然灾害影响和田赋混乱方面起到了显著效果，形成了解州和介休两大清丈范例，为其他地区提供了参考。

史料列表

史料1:

>光绪初年的“丁戌奇荒”进一步加剧山西田赋征收的混乱。自然灾害导致大规模人口逃亡与土地抛荒，土地产权发生剧烈变动，私占、倒换现象普遍。大量地方里甲书差亡失更替，田赋体系进一步崩溃。

——《张之洞全集》

史料2:

>解州清丈从光绪五年十月开始，到光绪七年五月结束，历时一年半多。当地认为，“清地先清粮，欲清粮先清户，三事必分先后，终究合为一事，递次详查”，在清查户地粮的基础上，进行点地、验契、丈地、抽丈、核算地粮、誊写工作，最终形成各村鱼鳞册与征粮簿。

——《解州丈清地粮里甲图说》

史料3:

>光绪五年的解州清丈办法创之于当地乡绅孙守恒，推广至山西阳曲、陕西朝邑等地，未受介休清丈办法的直接影响。

——《高崇基行之于介休而效，马丕瑶行之于解州而益效》

史料4:

>解州、阳曲等地清丈活动中，解州清丈成为可效仿的范例，而阳曲则是在解州清丈基础上进一步发展出的另一套清丈系统。

——《光绪解州志》

历史推理过程

自然灾害导致的大规模人口逃亡和土地抛荒加剧了田赋混乱（史料1）+解州通过清丈活动形成了详细准确的土地登记和征粮体系（史料2）->解州的清丈活动有效缓解了自然灾害造成的田赋混乱，形成了一个可效仿的范例（推论1）

解州清丈成为可效仿的范例对其他地区有指导意义（史料3）+阳曲在解州清丈的基础上进一步发展出另一套清丈系统（史料4）->解州和阳曲的清丈实践为其他地区提供了参考（推论2）

推论1+推论2->解州、阳曲等地的土地清丈在应对自然灾害影响和田赋混乱方面起到了显著效果，形成了解州和介休两大清丈范例，为其他地区提供了参考。（作者观点）

####

观点3

####

作者观点

清丈活动通过编制详细准确的鱼鳞册，缓解了田赋混乱问题，并帮助地方政府更好地管理土地和征收赋税。

史料列表

史料1:

>雍正以后，清丈扰民的认识逐渐成为主流。乾隆即位后采纳大学士朱轼的建议，“将丈量一事永行停止”。此后，一些省份虽有局部清丈活动，但多针对新垦荒地、滩地、沙田等类土地，官方组织的全面清丈基本停止，鱼鳞册失修现象十分普遍。

——《晚清山西土地清丈》

史料2:

>解州清丈从光绪五年十月开始，到光绪七年五月结束，历时一年半多。当地认为，“清地先清粮，欲清粮先清户，三事必分先后，终究合为一事，递次详查”，在清查户地粮的基础上，进行点地、验契、丈地、抽丈、核算地粮、誊写工作，最终形成各村鱼鳞册与征粮簿。

——《解州丈清地粮里甲图说》

史料3:

>清丈后，各村绘制地亩图。现存同治十一年介休《西段屯地理图》碑即是依据此次清丈结果绘制而来。

——《晚清山西土地清丈与鱼鳞册攷造》

历史推理过程

鱼鳞册失修导致田赋混乱（史料1）+ 解州建立了新的鱼鳞册系统，并通过清丈活动明确地籍和粮额（史料2）

->清丈活动帮助缓解了田赋混乱问题（推论1）

清丈后，各村绘制了地亩图，以确保地亩登记准确（史料3）+ 清丈活动中详细明确地籍和征粮簿的编制（史料2）

->清丈活动帮助地方政府更好地管理土地和征收赋税（推论2）

推论1+推论2->清丈活动通过编制详细准确的鱼鳞册，缓解了田赋混乱问题，并帮助地方政府更好地管理土地和征收赋税（作者观点）

####

观点4

####

作者观点

解州和介休的清丈方法各自发展，形成了各自独特且独立的清丈系统。

史料列表

史料1:

>光绪五年的解州清丈办法创之于当地乡绅孙守恒，推广至山西阳曲、陕西朝邑等地，未受介休清丈办法的直接影响。

——《高崇基行之于介休而效，马丕瑶行之于解州而益效》

史料2:

>解州、阳曲等地清丈活动中，解州清丈成为可效仿的范例，而阳曲则是在解州清丈基础上进一步发展出的另一套清丈系统。

——《光绪解州志》

历史推理过程

解州清丈办法创自当地乡绅孙守恒，推广至山西阳曲和陕西朝邑，但未受介休清丈办法的直接影响（史料1）+ 阳曲和解州清丈系统进一步发展，形成了独立的体系（史料2）->解州和介休的清丈方法各自发展，形成了各自独特且独立的清丈系统（作者观点）

#####

-Real Data-

#####

现在，根据下述论文，生成4个观点。此处，请允许我再强调一遍重要的注意事项，这对我的工作非常重要：

1. 每个观点的生成是互相独立的，各自须包含作者观点、史料列表和历史推理过程，但您应该使得4个选项都正确但各不相同，它们可以反映作者在不同事件、不同角度、不同分论点上的观点。
2. 每一个推理符”->“左侧的节点**至少有2个节点**，每一行仅有一个推理符”->“，禁止使用现有作者观点进行历史推理，禁止一个节点解释多个史料!!!
3. 不允许无中生有，不允许编造论文中不存在的观点和史料!!!

Generating False Option Candidates

-Goal-

历史是一门注重逻辑推理和严密论证的实证性的人文社会学科。遵循史料研究注意事项，根据给定史料和现有选项，生成普通人可能推导出的错误观点，并生成能够有效反驳该错误观点的历史推理过程。

-Knowledge-

史料研究注意事项：

1. 史料价值分类。第一手史料（直接史料）：原始资料，即历史人物、事件、现象发生的那个时代留下来的材料，特别是当事人和目击者提供的史料，直接反映了所研究对象的情况。第二手材料（间接史料）：距离那个时代较远的记录、转述，或者后人的研究成果，间接反映了研究对象的情况。
2. 对史料要合理利用，做到史论结合论由史出，有一分材料说一分话。
3. 搜集更多的史料，对多种史料进行甄选、辨别，多源互证。
4. 孤证不立：尽可能多地搜集史料。

-Constraints-

1. 基于基本逻辑和一定的历史背景知识，人类可以根据历史推理过程来反驳错误观点。
2. 证据链的每一节点对应一个史料，证据链必须至少包含2个以上的史料。
3. 优先使用一手史料，当一手史料用完之后，再考虑使用二手史料。
4. 您生成的错误观点，不允许与给定的现有观点重复，不允许在历史推理过程中使用现有观点中的信息。
5. 您生成的错误观点，需要具有较高的辨别难度，需要根据多个史料信息进行联合推理后，才能进行反驳和排除。
6. 每次回复中，您须严格遵循下述输出格式，不要回复无关内容。YOU MUST FOLLOW THE OUTPUT FORMAT BELOW.

-Input Format-

史料列表
史料列表

现有观点

已生成的现有观点，最多不超过3个，最少为1个

-Output Format-

错误观点

给定史料列表，生成一般人容易错误推导出的观点，要求如下：

- 该错误观点至少需要用2个给定史料才能够判别正误。
- 尽可能提高该错误观点的判断难度!!!
- 不允许使用“绝对性”的表述，例如“完全”、“任何”、“绝对”、“唯一”、“任何”等，因为这样很可能降低错误选项的辨别难度!!!
- 减少使用否定性表述，例如“并没有”、“并未”、“忽略”、“很少”等，因为这样很可能降低错误选项的辨别难度!!!
- 该错误观点不用该与现有观点重复或语义相似。

历史推理过程

根据给定史料，至少使用2个史料来反驳[1]中的错误观点，给出详细的历史推理过程。历史推理过程包含单步推理或多步推理两种可能情况：

1. 单步推理：每一行仅允许有一个推理符“->”，该推理符左侧**至少有2个节点**，最多不超过4个，且均为史料；该推理符右侧**仅有1个节点**，且必然为作者观点；
史料1的解释（史料1）+ 史料2的解释（史料2）+ ... ->反驳错误观点的解释（反驳错误观点）
2. 多步推理：每一行仅允许有一个推理符“->”，每一个推理符左侧的节点**至少有2个节点**，最多不超过4个，这些节点可能是史料，也可能是已有的推论；推理符右侧**仅有1个节点**，非最后一行的右侧节点必然是新的推论，最后一个右侧节点必然为作者观点；
史料1的解释（史料1）+ 史料2的解释（史料2）->推论1的内容（推论1）
史料3的解释（史料3）+ 史料4的解释（史料4）->推论2的内容（推论2）
推论1 + 推论2 ->反驳错误观点的解释（反驳错误观点）

推理约束：

- a. 您应当确保，您每一行历史推理过程都要符合人类常识和基本逻辑。
- b. “->”左侧的每个史料，都应该对推导出“->”右侧的推论有重要作用。
- c. 推论从1开始计数，新的推论在原有最大推论序号上+1，推论的序号不允许重复和跳过。
- d. 禁止使用现有观点进行任何历史推理！！
- e. 基于史料生成的推论，至少需要被作为输入使用一次，不允许仅生成而不被使用。
- f. 禁止根据单个史料、或者单个推论，来推理出新的推论或作者观点：
- g. 禁止一个节点解释多个史料，如：
 - 这是错误的！“史料m和史料n的解释（史料m和史料n）+...->新的节点“，这是错误的！”
 - 这是错误的！“史料m和史料n的解释（史料m与史料n）+...->新的节点“，这是错误的！”
 - 这是错误的！“史料m和史料n的解释（史料m，史料n）+...->新的节点“，这是错误的！”
 - 这是错误的！“史料m和史料n的解释（史料m+史料n）+...->新的节点“，这是错误的！”
- h. 每一行仅有一个推理符“->“！

#####

-Examples-

#####

Example 1:

#####

Input:

史料列表

>自淮北置巡院十三，曰扬州、陈许、汴州、庐寿、白沙、淮西、甬桥、浙西、宋州、泗州、岭南、兖郛、郑滑，捕私盐者，奸盗为之衰息。

——《新唐书·食货志》

史料2:

>晏始以盐利为漕佣，自江淮至涇桥，率十万斛佣七千缗，补纲吏督之。不发丁男，不劳郡县，盖自古未有之也。自此岁运米数（千）〔十〕万石，自淮北列置巡院，搜择能吏以主之，广牢盆以来商贾。

——《旧唐书·食货志》

史料3:

>臣自到州，见租庸等诸使文牒，令征前件钱物送纳。臣当州被西原贼屠陷.....伏望天恩，自州未破已前，百姓久负租税，及租庸等使所有征率和市杂物，一切豁免。自州破以后，除正租正庸，及准格式合进奉征纳者，请据见在户征送。其余科率，并请豁免。

——《元次山集》

史料4:

>诸道各置知院官，每旬月，具州县雨雪丰歉之状白使司，丰则贵余，歉则贱糶，或以谷易杂货供官用，及于丰处卖之。知院官始见不稔之端，先申，至某月须如千蠲免，某月须如千救助，及期，晏不俟州县申请，即奏行之，应民之急，未尝失时，不待其困弊、流亡、饿殍，然后赈之也。

——司马光《资治通鉴》

史料5:

>东自淮阴，西临蒲坂，亘三千里，屯戍相望。中军皆鼎司元侯，贱卒亦仪同青紫。每云食半菽，又云无拱紆，漕所至，船到便留，即非单车使折简书所能制矣。

——《刘晏集》

史料6:

>寇戎以来，积有年岁。征求数广，凋弊转深。自今已后，除正租税及正敕并度支符外，余一切不在征科限内。

——《广德二年南郊赦》

史料7:

>刘晏上书元载，强藩挟制、单车使无力应对。

——刘晏《上元载书》

史料8:

>无何，迁殿中侍御史，充荆、襄转运留后。其时，节使有恃兵柄，征赋多愆，诏征不入者。君体柔谦以尽祗

敬，陈大义以奖忠烈。利害相悬，狂昏乃革，或毕词而飙发泉布，或不信宿而骏奔朝谒。其释疑开物，有如此者。他年又加洪、潭、黔、广盐铁租庸等使。爰领六道，凡经五年，物无浮窳，人不怨，上司取给，伊君是赖。

——《杨墓志》

史料9:

> 唐廷一面要与沿线藩镇合作，如先后与河南副元帅李光弼计会开决汴水、委河南副元帅王缙“各与本道节度计会商量，夹河两岸，每两驿置防援三百人”以防盗贼，一面又要尽量加强自己在漕运上的主动性。

——《旧唐书》

现有观点

A. 代宗时期的道租庸使在税收和转运上的职能出现了分解和重新组合，逐渐演变为诸道巡院与转运留后，这些机构在不同方面辅助了国家赋税体系的重建。

B. “刘晏置十三巡院”的记载包含了后代设立的巡院，因此在历史真实性上需要谨慎对待，但其中汴路运河沿线的巡院是为漕运提供资金而设立的，符合历史背景。

C. 广德二年（764）刘晏开设汴河巡院的主要目的是通过盐专卖来资助漕运，并制定了一系列与州县合作的措施以保障漕运顺畅。

#####

Output:

错误观点

刘晏的盐专卖改革推进过程中，强藩问题得到了有效解决，从而确保了漕运的顺畅。

历史推理过程

刘晏上书元载指明强藩问题及单车使无力应对（史料7）+ 唐廷需要与沿线藩镇合作以保障漕运顺畅（史料9）

-> 强藩问题在刘晏推进盐专卖改革的过程中并未完全解决，而是需要通过不断与沿线藩镇的合作来维持漕运的顺畅（反驳错误观点）

#####

Example 2:

#####

Input:

史料列表

史料1:

> 夫以孔子之圣，平生所志，惟在东周，生今反古，深用为戒。老不得行其道，犹修《春秋》以存周典，此岂以周之法独善于前代哉？

——张居正之言

史料2:

> 朕不敢不勉行法祖，然凡事尚赖先生辅导。

——《明实录》

史料3:

> 近年以来，好事者喜于纷更，建议者鲜谄国体，条例纷纭……耳目淆惑，莫知适从，我祖宗之良法美意几于沦失矣。

——张居正《淮安府志》

史料4:

> 方今国家要务，惟在遵守祖宗旧制，不必纷纷更改。

——《明实录》

史料5:

> 我国家之兴，始皇祖高皇帝也。中定艰难，则我皇祖文皇帝也。二圣同创大业，功德并焉，宜称祖号。

——《明世宗实录》

史料6:

>正德年间，政体紊乱。至世宗皇帝，以聪明至圣入继大统，将以前敝政一切改正，以复我祖宗之旧，正今日所当遵守者。

——《张居正全集》

史料7:

>我圣祖以天之心为心，故能创造洪业，传之皇上。在皇上今以圣祖之心为心，乃能永保洪业，传之无穷。

——《帝鉴图说》

史料8:

>夫法制无常，近民为要，古今异势，便俗为宜。孟子曰：“遵先王之法而过者，未之有也。”此欲法先王矣。荀卿曰：“略法先王而足乱世术，不知法后王而一制度，是俗儒者也。”此欲法后王矣。两者互异，而荀为近焉。何也？法无古今，惟其时之所宜，与民之所安耳。时宜之，民安之，虽庸众之所建立，不可废也；戾于时，拂于民，虽圣哲之所创造，可无从也。

——张居正《辛未会试程策》

现有观点

A. 张居正认为法度应根据时代变化和民众需求而变，不应盲目遵古，更应实行因时制宜的政策。

#####

Output:

错误观点

张居正“法后王”思想的核心主要是以明代“二祖”为榜样，注重历史传承与现实政治的融合，通过实际执行合乎时宜的祖宗法度以实现良治。

历史推理过程

张居正反对“法先王”，认为法制应根据现实需求灵活变通（史料8）+ 张居正推崇明太祖和成祖的功绩（史料5）->张居正主张“法后王”，如明太祖和明成祖，认为法度应适应时代需求，且应以安民为要（推论1）

推论1+张居正高度评价明世宗的改革举措（史料6）->张居正认为注重历史传承与现实政治的融合，通过实际执行合乎时宜的祖宗法度以实现良治，观点的后半句是正确的；但“法后王”思想的核心是不仅是以明代“二祖”为榜样，也需要以明世宗为榜样，如果无视张居正对明世宗“中兴”地位的肯定，那就不可能客观地把握张居正“法后王”的思想要义（反驳错误观点）

#####

-Real Data-

#####

现在，根据下述史料列表和现有观点，生成一个错误观点以及反驳该错误观点的历史推理过程。记住，每一个推理符“->“左侧的节点**至少有2个节点**，每一行仅有一个推理符”->“，禁止使用现有作者观点进行历史推理，禁止一个节点解释多个史料，您现在生成的观点必须与给定的现有观点不同（比如不同角度、不同意见）!!! 这对我的工作非常重要。

Verifying Generated True Options

-Goal-

历史是一门注重逻辑推理和严密论证的实证性的人文社会科学。遵循史料研究注意事项，根据给定史料，检查当前观点以及对应的历史推理过程是否正确。

-Knowledge-

史料研究注意事项：

1. 史料价值分类。第一手史料（直接史料）：原始资料，即历史人物、事件、现象发生的那个时代留下来的材料，特别是当事人和目击者提供的史料，直接反映了所研究对象的情况。第二手材料（间接史料）：距离那个时代较远的记录、转述，或者后人的研究成果，间接反映了研究对象的情况。
2. 对史料要合理利用，做到史论结合论由史出，有一分材料说一分话。
3. 搜集更多的史料，对多种史料进行甄选、辨别，多源互证。
4. 孤证不立：尽可能多地搜集史料。

-Constraints-

1. 基于基本逻辑和一定的历史背景知识，人类可以根据历史推理过程来推理出当前观点。
2. 证据链的每一节点对应一个史料，证据链必须至少包含2个以上的史料。
3. 优先使用一手史料，当一手史料用完之后，再考虑使用二手史料。
4. 每次回复中，您须严格遵循下述输出格式，不要回复无关内容。YOU MUST FOLLOW THE OUTPUT FORMAT BELOW.
5. 以下是**历史推理过程的形式要求**，历史推理过程包含单步推理或多步推理两种可能情况：
 - a. 单步推理：每一行仅允许有一个推理符“->”，该推理符左侧**至少有2个节点**，最多不超过4个，且均为史料；该推理符右侧**仅有1个节点**，且必然为作者观点；
史料1的解释（史料1）+ 史料2的解释（史料2）+ ... ->作者观点的内容（作者观点）
 - b. 多步推理：每一行仅允许有一个推理符“->”，每一个推理符左侧的节点**至少有2个节点**，最多不超过4个，这些节点可能是史料，也可能是已有的推论；推理符右侧**仅有1个节点**，非最后一行的右侧节点必然是新的推论，最后一个右侧节点必然为作者观点；
史料1的解释（史料1）+ 史料2的解释（史料2）->推论1的内容（推论1）
史料3的解释（史料3）+ 史料4的解释（史料4）->推论2的内容（推论2）
推论1 + 推论2 ->作者观点的内容（作者观点）

推理约束：

- a. 您应当确保，您每一行历史推理过程都要符合人类常识和基本逻辑。
- b. “->”左侧的每个史料，都应该对推导出“->”右侧的推论有重要作用。
- c. 推论从1开始计数，新的推论在原有最大推论序号上+1，推论的序号不允许重复和跳过。
- d. 基于史料生成的推论，至少需要被作为输入使用一次，不允许仅生成而不被使用。
- e. 禁止根据单个史料、或者单个推论，来推理出新的推论或作者观点。
- f. 禁止一次性解释多个史料，如：
 - 这是错误的！“史料m和史料n的解释（史料m和史料n）->新的节点“，这是错误的！
 - 这是错误的！“史料m和史料n的解释（史料m与史料n）->新的节点“，这是错误的！
 - 这是错误的！“史料m和史料n的解释（史料m，史料n）->新的节点“，这是错误的！
 - 这是错误的！“史料m和史料n的解释（史料m+史料n）->新的节点“，这是错误的！
- h. 每一行仅有一个推理符“->”！

-Input Format-

史料列表

多个一手或二手的史料，但注意，并非所有选项都与当前观点有关

当前观点（正确观点）

当前观点

当前观点的历史推理过程

历史推理过程

-Output Format-

分析

分析过程应该尽可能简洁，需要考虑的角度包括：

- 根据史料，判断当前观点是否正确
- 当前观点在大多数情况是可靠的，你不应该轻易修改他们，**仅仅在基于现有史料不足以推理出当前观点的时候**，可以对该观点进行最小幅度的修改和语义收窄（但不应与原意冲突），使其基于现有史料的推理可以推理出来（注意修改相应的历史推理过程）。
- 请分析历史推理过程中的每一步推理，是否符合人类逻辑、历史常识和推理约束。

修改后的观点（必须仍然是正确观点）

要求：

- 如果当前观点正确，则回复“无需修改”。
- 如果当前观点错误或基于现有史料不足以推理出来，那么您需要对当前观点进行修改。注意，尽可能只做最小程度的修改，而不是大幅修改甚至是重新生成。

修改后的历史推理过程

要求：

- 您修改后的史料必须仍然严格遵循**历史推理过程的形式要求**，如果原本的历史推理过程不遵循**历史推理过程的形式要求**，您也需要相应地进行修改，使其遵循要求!!!
- 如果当前历史推理过程存在问题，那么修改该历史推理过程，使其符合人类逻辑和历史常识，且基于该推理过程能够验证当前观点是正确的。注意，尽可能只做最小程度的修改，而不是大幅修改甚至是重新生成。
- 如果当前历史推理过程已符合人类逻辑、历史常识和推理约束，则回复“无需修改”。
- 最终节点必须标注“（作者观点）”

#####

-Examples-

#####

Example 1:

#####

Input:

史料列表

史料1:

>幽、厉之后，周室微，陪臣执政，史不记时，君不告朔，故疇人子弟分散，或在诸夏，或在夷狄。

——《史记·历书》

史料2:

>屈原于《离骚》中言：“世系长于天下，为政有户，皆学之至。”传世文献表明，疇官的形成具有深远的历史背景，涉及世官制瓦解后的职业流动和专业人才的世袭传承。

——《屈原研究》

史料3:

《史记》记载：

>至周室之卜官……及孝惠享国日少，吕后女主，孝文、孝景因袭掌故，未遑讲试，虽父子疇官，世世相传，其精微深妙，多所遗失。

——《史记·龟策列传》

史料4:

《文献通考·兵考》记载：

>凡民年二十三，附之疇官。给郡县一月而谓更卒，复给中都一岁谓正卒，复屯边一岁谓戍卒。

——《文献通考》

史料5:

《岳麓秦简（柒）》记载：

>十九年八月辛丑，丞相请：恒以傅时试疇司御、医、鼓人、执剂、骚医之新傅，不中试者，夺疇令戍新地三岁，日备，勿令复疇。

——《岳麓秦简（柒）》

史料6:

章太炎认为：

>周官三百七十有余品，约其文辞，其凡目在疇人世官。

——章太炎《周礼》

史料7:

>《明实录》记载：“史、卜、祝子年十七岁学，史、卜、祝学童学三岁……学俱将诣大史、大卜、大祝，郡史学童诣其守，皆会八月朔日试之。”

——《明实录》

史料8:

《岳麓书院藏秦简（肆）》记载：

>房学吹和讴隶臣妾可以其音乐技能免为“学子”，其中的讴隶臣妾在“学吹讴”时称“学子”。

——《岳麓书院藏秦简》

史料9:

>疇尸、菑御、杜主乐皆五更，属大祝。祝年盈六十者，十二更，践更大祝。

——《张家山汉简》

史料10:

>张家山汉简《二年律令·傅律》规定不同爵位之人，其子应在不同年龄段傅籍，第364—366号简提到：不更以下子年廿岁……疇官各从其父疇，有学师者学之。当傅，高不盈六尺二寸以下，及天鸟，皆以为罢癯。

——《张家山汉简》

史料11:

《二年律令·史律》记载:

>史、卜受调书太史、太卜而逋留，及擅不视事盈三月，斥勿以为史、卜。

——《张家山汉简》

史料12:

>清代阮元编撰《疇人传》，汇集上古至嘉庆年间两百多位专精天文、历法和算学的学者，正是对这一传统认识的赓续。受传世文献及其旧注所限，今人对疇官的认识一般只限于史卜星历之官。

——《疇人传》

当前观点（正确观点）

秦汉时期的疇官不仅包括传统的史、卜、祝、司御等职业，还涵盖了一系列其他的专业技能职业，疇官制度是在周代世官制瓦解后的延续，体现了专业技能在世代传承和国家管理中的重要性。

当前观点的历史推理过程

周代世官制的背景和疇官的传承（史料6）+周代卜官等传统职业的传承至秦汉（史料1）->疇官制度是周代世官制的延续（推论1）

疇人为执掌专业知识的职官并有世代相传的特点（史料1）+秦汉时期实际存在不止史、卜等传统职业，还包括其他如疇尸、菑御、杜主乐等职业（史料9）->秦汉时期的疇官范围广泛，涵盖了多个专业技能职业（推论2）

秦汉时期疇官必须通过官方考核，并且涉及许多职业如疇司御、医、鼓人等（史料5）+推论1+推论2->秦汉时期的疇官不仅包括传统的史、卜等职业，还涵盖了其他专业技能职业，反映了专业技能在国家管理中的重要性，并延续了周代的世官制（作者观点）

#####

Output:

分析

- 观点：当前观点是正确的，无需修改。

- 历史推理过程：符合人类逻辑、历史常识和推理约束，能够基于该推理过程顺利推理出当前观点是正确的，无需修改

修改后的观点（必须仍然是正确观点）

无需修改

修改后的历史推理过程

无需修改

#####

Example 2:

#####

Input:

史料列表

史料1:

>根据调查，景德镇传统民窑的生产技术、产品质量和生产成本在一定程度上受到了御窑的影响与带动。

——王世襄《清代匠作则例汇编》

史料2:

>影响因素的敏感度分析中，价格和成本对景德镇瓷业实际产值的敏感度最高，特别在晚清时期。清朝末期所有要素的波动较大，价格的波动尤为显著。

——《景德镇陶瓷概况》

史料3:

>清后期瓷器包装物料价格大幅上涨,这解释了清代瓷业产值高峰出现在光绪朝的原因。

——《清宫瓷器档案全集》

史料4:

>咸丰至同治初年是景德镇瓷器烧造数量的低谷阶段,但光绪朝后,生产数量相比前期大幅度减少,且清末景德镇单件瓷器的生产成本大幅提高,光绪时期单件瓷器的生产成本约为雍乾时期的6倍左右。

——《清宫瓷器档案全集》

史料5:

>瓷器生产受到当地的天气影响,如果天气寒冷,泥土干冻,就无法进行瓷器生产,开工时间会延后。清代御窑生产过程中有一标准环节名为“大开工祭祀”,即开工之前的祭祀活动。

——《清宫瓷器档案全集》

史料6:

>雍正十三年年的供词中明确提到当时的御窑生产已经转为厂内制作坯胎民窑进行烧造,说明御窑与民窑的生产是有紧密联系的。

——舒善《供词》

史料7:

>清宫瓷器档案所载从雍正四年至宣统三年的御窑瓷器详细记录,这些记录包括了各朝各年的御窑生产情况、成本与燃料费用等详细数据。

——铁源、李国荣《清宫瓷器档案全集》

史料8:

>高岭土对瓷器的生产至关重要,清代开采和使用的高岭土主要来源地有浮梁东埠高岭村、黄坛大洲村和九江星子;其中东埠土被替代,增加了开采成本和运输成本,影响了清代瓷器的产值。

——《景德镇陶瓷史》

史料9:

>道光二十九年,景德镇的开工祭祀延迟了八天,冬季平均气温每降低1摄氏度,大开工祭祀的时间就会晚8.84天。

——《江西通志》

当前观点(正确观点)

御窑生产的记录与分析不仅有助于理解清代景德镇瓷业的发展情况,还能提供一种有效的方法来推算清代民窑的生产情况,弥补了民窑直接数据缺乏的局限。

当前观点的历史推理过程

清宫瓷器档案详细记录了御窑的生产情况及成本(史料7)+御窑生产的坯胎被交由民窑烧造,表明御窑与民窑有紧密联系(史料6)->御窑的生产记录有助于理解清代景德镇瓷业的发展情况(推论1)

御窑技术、质量和成本直接影响民窑的生产情况(史料1)+天气等外部因素影响御窑的生产,也同样影响民窑的开工时间和生产情况(史料5)->御窑生产记录可以用于推算清代民窑的生产情况,提供了一种弥补民窑数据缺乏局限的方法(作者观点)

#####

Output:

分析

- 观点:当前观点是正确的,无需修改。

- 历史推理过程:该历史推理过程存在一定问题,推论1仅生成而不被使用,不符合推理约束。其次,历史推理过程中的最终节点(作者观点)并未将当前观点表述完整,仅表述了后半部分。

修改后的观点(必须仍然是正确观点)

无需修改

修改后的历史推理过程

清宫瓷器档案详细记录了御窑的生产情况及成本（史料7）+ 御窑生产的坯胎被交由民窑烧造，表明御窑与民窑有紧密联系（史料6）->御窑的生产记录有助于理解清代景德镇瓷业的发展情况（推论1）

御窑技术、质量和成本直接影响民窑的生产情况（史料1）+ 天气等外部因素影响御窑的生产，也同样影响民窑的开工时间和生产情况（史料5）->御窑生产记录可以用于推算清代民窑的生产情况，提供了一种弥补民窑数据缺乏局限的方法（推论2）

推论1+推论2->御窑生产的记录与分析不仅有助于理解清代景德镇瓷业的发展情况，还能提供一种有效的方法来推算清代民窑的生产情况，弥补了民窑直接数据缺乏的局限。（作者观点）

#####

Example 3:

#####

Input:

史料列表

史料1:

>我圣祖以天之心为心，故能创造洪业，传之皇上。在皇上今以圣祖之心为心，乃能永保洪业，传之无穷。

——《明实录》

史料2:

>明世宗起自藩服，光履帝位，实上天之所眷与，四海之所归心，非偶然者。故能覲扬前休，廓清区宇，制礼作乐，开四十五年中兴之业，而垂德泽于无穷，近代帝王盖鲜俚云。

——《明世宗实录》

史料3:

>孟子曰：“遵先王之法而过者，未之有也。”此欲法先王矣。荀卿曰：“略法先王而足乱世术，不知法后王而一制度，是俗儒者也。”此欲法后王矣。两者互异，而荀为近焉。何也？法无古今，惟其时之所宜，与民之所安耳。时宜之，民安之，虽庸众之所建立，不可废也；戾于时，拂于民，虽圣哲之所创造，可无从也。

——《辛未会试程策》

史料4:

>我国家之兴，始皇祖高皇帝也。中定艰难，则我皇祖文皇帝也。二圣同创大业，功德并焉，宜称祖号。

——《明世宗实录》

史料5:

>法当宜民，政以人举，民苟宜之，何分南北？

——《张居正集》

史料6:

>近年以来，好事者喜于纷更，建议者鲜谄固体，条例纷纭.....耳目淆惑，莫知适从，我祖宗之良法美意几于沦失矣。

——张居正《淮安府志》

史料7:

明世宗即位时的诏书明确指出：

>祖宗成法，至精至备，所当万世遵守。近年以来，有司不考宪度，往往自作聪明，任意更变。其有称为祖宗成法者，又多迁移出入，殊非祖宗立法本意，致令事体纷纭，军民惶惑，岂成治理？今后内外大小衙门官，务要仰求祖宗之意，明考成法，一一遵行，违者以变乱成法论。

——《明实录》

史料8:

>夫以孔子之圣，平生所志，惟在东周，生今反古，深用为戒。老不得行其道，犹修《春秋》以存周典，此岂以周之法独善于前代哉？盖为下之礼宜尔也。今世俗皆曰“愿学孔子”，乃不务遵祖宗之典，以服官寡过。而好言上古久远之事，以异趋为高；动循衰世苟且之政，以徇情贾誉。此岂圣人所谓“为下不倍”哉！恶在其为遵孔氏也。

——张居正《辛未会试程策》

当前观点（正确观点）

张居正认为嘉靖时期的明世宗是明朝的“中兴”之主，他的革新之法值得后世效法，与明太祖和明成祖并列为后王。

当前观点的历史推理过程

法当宜民，政以人举，常规应因时而变（史料5）+ 中定艰难、制礼作乐并垂德泽于无穷的明世宗被被誉为中兴之君（史料2）->明世宗的中兴地位得到了部分承认（推论1）

张居正强调要尊崇孔子的礼法且批评当代士风（史料8）+ 明世宗在诏书中提到要遵行祖宗成法（史料7）->张居正认可并重视明世宗的某些革新之法（推论2）

推论1+推论2->张居正认为嘉靖时期的明世宗是明朝的“中兴”之主，他的革新之法值得后世效法，与明太祖和明成祖并列为后王。（作者观点）

#####

Output:

分析

- 观点：基于现有史料不足以推理出当前观点，因此对该观点进行最小幅度的修改和语义收窄（但不应与原意冲突），使其基于现有史料的推理可以推理出来。
- 历史推理过程：对当前观点进行修改后，其历史推理过程中的相关部分也需要进行修改。

修改后的观点（必须仍然是正确观点）

嘉靖时期的明世宗是明朝的“中兴”之主，张居正认为他的革新之法值得后世效法。

修改后的历史推理过程

法当宜民，政以人举，公司常规应因时而变（史料5）+ 中定艰难、制礼作乐并垂德泽于无穷的明世宗被被誉为中兴之君（史料2）->明世宗的中兴地位得到了部分承认（推论1）

张居正强调遵循孔子的做法，选择符合时宜的古法（史料8）+ 明世宗在诏书中提到要遵行祖宗成法（史料7）->张居正认可并重视明世宗的某些革新之法（推论2）

推论1+推论2->嘉靖时期的明世宗是明朝的“中兴”之主，张居正很可能会认为他的革新之法值得后世效法。（作者观点）

#####

-Real Data-

#####

记住，每一个推理符“->“左侧的节点**至少有2个节点**”，每一行仅有一个推理符“->“，禁止一个节点解释多个史料!!! 这对我的工作非常重要。

Verifying Generated False Options

-Goal-

历史是一门注重逻辑推理和严密论证的实证性的人文社会学科。遵循史料研究注意事项，根据给定史料，检查当前的错误观点是否有足够的难度，以及反驳该错误观点的历史推理过程是否正确。

-Knowledge-

史料研究注意事项：

1. 史料价值分类。第一手史料（直接史料）：原始资料，即历史人物、事件、现象发生的那个时代留下来的材料，特别是当事人和目击者提供的史料，直接反映了所研究对象的情况。第二手材料（间接史料）：距离那个时代较远的记录、转述，或者后人的研究成果，间接反映了研究对象的情况。
2. 对史料要合理利用，做到史论结合论由史出，有一分材料说一分话。
3. 搜集更多的史料，对多种史料进行甄选、辨别，多源互证。
4. 孤证不立；尽可能多地搜集史料。

-Constraints-

1. 基于基本逻辑和一定的历史背景知识，人类可以根据历史推理过程来反驳错误观点。

2. 证据链的每一节点对应一个史料，证据链必须至少包含2个以上的史料。

3. 优先使用一手史料，当一手史料用完之后，再考虑使用二手史料。

4. 您修改后的错误观点，需要具有较高的辨别难度，需要根据多个史料信息进行联合推理后，才能进行反驳和排除。

5. 每次回复中，您须严格遵循下述输出格式，不要回复无关内容。YOU MUST FOLLOW THE OUTPUT

FORMAT BELOW.

6. 以下是**历史推理过程的形式要求**，历史推理过程包含单步推理或多步推理两种可能情况：

a. 单步推理：每一行仅允许有一个推理符“->”，该推理符左侧**至少有2个节点**，最多不超过4个，且均为史料；该推理符右侧**仅有1个节点**，且必然为作者观点；
史料1的解释（史料1）+ 史料2的解释（史料2）+ ... ->反驳错误观点的解释（反驳错误观点）

b. 多步推理：每一行仅允许有一个推理符“->”，每一个推理符左侧的节点**至少有2个节点**，最多不超过4个，这些节点可能是史料，也可能是已有的推论；推理符右侧**仅有1个节点**，非最后一行的右侧节点必然是新的推论，最后一个右侧节点必然为作者观点；

史料1的解释（史料1）+ 史料2的解释（史料2）->推论1的内容（推论1）

史料3的解释（史料3）+ 史料4的解释（史料4）->推论2的内容（推论2）

推论1+推论2->反驳错误观点的解释（反驳错误观点）

推理约束：

a. 您应当确保，您每一行历史推理过程都要符合人类常识和基本逻辑。

b. “->”左侧的每个史料，都应该对推导出“->”右侧的推论有重要作用。

c. 推论从1开始计数，新的推论在原有最大推论序号上+1，推论的序号不允许重复和跳过。

d. 基于史料生成的推论，至少需要被作为输入使用一次，不允许仅生成而不被使用。

e. 禁止根据单个史料、或者单个推论，来推理出新的推论或作者观点：

f. 禁止一次性解释多个史料，如：

- 这是错误的！“史料m和史料n的解释（史料m和史料n）->新的节点“，这是错误的！

- 这是错误的！“史料m和史料n的解释（史料m与史料n）->新的节点“，这是错误的！

- 这是错误的！“史料m和史料n的解释（史料m，史料n）->新的节点“，这是错误的！

- 这是错误的！“史料m和史料n的解释（史料m+史料n）->新的节点“，这是错误的！

h. 每一行仅有一个推理符“->”！

-Input Format-

史料列表

多个一手或二手的史料，但注意，并非所有选项都与当前观点有关

当前观点（有辨别难度的错误观点）

当前观点

反驳当前观点的历史推理过程

历史推理过程

-Output Format-

分析

分析过程应该尽可能简洁，需要考虑的角度包括：

- 根据史料，判断当前观点是否错误，以及是否具有较高的辨别难度。

- 当前观点如果是正确的，那需要将其修改为错误的，且使其具备较高的辨别难度（即至少需要根据2个史料进行联合推理后，才能进行反驳和排除）。

- 当前观点如果是错误的，那么应考虑其是否具备比较高的判断难度（即至少需要根据2个史料进行联合推理后，才能进行反驳和排除），不允许使用“绝对性”的表述，例如“完全”、“任何”、“绝对”、“唯一”、“任何”等，因为这样很可能降低错误选项的辨别难度！！

- 请分析历史推理过程中的每一步推理，是否都符合人类逻辑、历史常识和推理约束。

修改后的观点（必须仍然是有辨别难度的错误观点）

要求：

- 如果当前观点是错误的，且具备较高的辨别难度，则回复“无需修改”。

- 当前观点如果是错误的，但辨别难度较低，那么尝试在不影响历史推理过程准确性的情况下，提高其辨别难度，不允许使用“绝对性”的表述，例如“完全”、“任何”、“绝对”、“唯一”、“任何”等，因为这样很可能降低错误选项的辨别难度！！减少使用否定性表述（“并没有”、“并未”、“忽略”、“很少”等），因为这样很可能降低错误选项的辨别难度！！

修改后的历史推理过程

要求:

- 您修改后的史料必须仍然严格遵循**历史推理过程的形式要求**，如果原本的历史推理过程不遵循**历史推理过程的形式要求**，您也需要相应地进行修改，使其遵循要求!!!
- 如果当前历史推理过程存在问题，那么修改该历史推理过程，使其符合人类逻辑和历史常识，且能用于反驳当前错误观点。注意，尽可能只做最小程度的修改，而不是大幅修改甚至是重新生成。
- 如果当前历史推理过程已符合人类逻辑、历史常识和推理约束，则回复“无需修改”。
- 最终节点必须标注“(反驳错误观点)”。

#####

-Examples-

#####

Example 1:

#####

Input:

史料列表

史料1

>查得国初原无边商、内商名色。自边商难于守支，故卖引于内商。内商难于报中，故买引于边商。一专报中，一专守支。

——《清理盐法疏》

史料2

>成化十九年颇减存积之数，常股七分，而存积三分。然商人乐有见盐，报中存积者争至，遂仍增至六分。淮、浙盐犹不能给，乃配支长芦、山东以给之。一人兼支数处，道远不及亲赴，边商辄贸引于近地富人。自是有边商、内商之分。

——《明史·食货志》

史料3

>成化年间，每引派纳折银三钱五分，递加至四钱二分。当此之时，每一边商引盐派定一运司，未有搭派别处之例，故商人亲中亲支，随支随掣，淮南岁掣八单，淮北四单，引盐流通，绝无阻格，亦无内商、边商之别。

——董光裕《议禁掣河盐》

史料4

>隆庆元年（1567）巡盐御史苏朝宗奏曰：“国初边商亲自支盐，至仪、淮二所掣卖。其后困于余盐，将河盐堆置淮扬，存积渐多，不暇守候，乃分拨引目，鬻之居民。故内商坐致富饶，而边商奔走益困。”

——《明实录》

史料5

>嘉靖三十二年，工本盐加增边商负担，商人不堪其苦，纷纷将盐引售于内商，遂使内商日益富有。

——《明史》

当前观点（有辨别难度的错误观点）

边商与内商的分化在成化年间就已完全实现，嘉靖时期的盐政改革仅是对此分化的延续和确认。

反驳当前观点的历史推理过程

成化年间未实现内商与边商的分化（史料3）+ 嘉靖时期（隆庆元年前）完成了内商与边商的分化（史料4）

->边商与内商的分化在成化年间就已完全实现不成立，预设观点错误（反驳错误观点）

#####

Output:

分析

- 观点：当前观点是错误的，也具备较高的判断难度，因此无需修改。

- 历史推理过程：符合人类逻辑、历史常识和推理约束，能够基于该历史推理过程可以反驳当前错误观点，无需修改。

修改后的观点（必须仍然是有辨别难度的错误观点）

无需修改

修改后的历史推理过程

无需修改

#####

Example 2:

#####

Input:

史料列表

史料1:

>田归道累迁左金吾将军、司膳卿,兼押千骑。未几,除尚方监,加银青光禄大夫。转殿中监,仍令依旧押千骑,宿卫于玄武门。

——《新唐书·田归道传》

史料2:

>《旧唐书·职官三》详细罗列了唐代武职事官的名号,分布于南衙十六卫,地方都督府、都护府、折冲府、镇戍的大将军、将军、中郎将及折冲都尉、果毅都尉等,品级森严、职事明确,都是名实一致的实职武官。

——张明《唐代武职事官阶官化》

史料3:

>武职事官检校、押领北衙禁军和充当行军军职这两种情况具有更为重要的影响。此处以阿史那忠为例,略做阐述。(永徽中)为左武卫大将军,寻迁右骁卫大将军。属兴师讨碣,以公为使持节长岑道行军大总管……(契丹)近侵并服,外结乌夷。公回师诛翦,应机殄灭,虏获万计,三军无私,蒙赏缣帛,仍于羽林军检校……总章元年,吐蕃入寇,拜使持节青海道行军大总管。

——《阿史那忠墓志》

史料4:

>唐太宗时期,武职事官首任为果毅都尉,逐步晋升至折冲都尉、中郎将等,这一过程较为顺畅。

——《高钦德基志》

史料5:

>咸亨元年,以左卫翊卫擢入羽林军……文明元年,奉敕简充引驾,仍每番明堂检校。载初元年,授左卫开方府左果毅都尉。其年九月,制加游击将军。如意元年,奉敕于玉门镇,经二周讨贼,再立功……圣历元年,除斜谷府折冲,奉敕于白草军防御。至军,又为左军总管……圣历二年,奉敕镇安西。至镇,又于拨换城守捉……神龙元年,制除廓州刺史,兼知积石军事……景龙二年九月七日,薨于积石军之公馆。

——《崔思忠墓志》

史料6:

>出身游击,制授将军,仍检校祥凤苑廐……寻以恪勤务,忠慎奉公。掌驥以呈材,效鷹而处事。迁宁远将军、守左豹韬卫洛府左果毅都尉,仍令长上,兼检校祥凤苑廐如故。而七萃之容载肃,五戎之寄逾殷。

——《索礼墓志》

史料7:

>安史之乱被评为大将军者五百余人,中郎将者二千余人,表明唐廷大量赏赐高级武职事官以激励军心。

——《旧唐书》

史料8:

>唐代马政系统较早地推行了使职差遣制,“马政系统的使职差遣制开始于马匹生产部门,是由当时严重缺马的现实决定的,并终于以彻底取代原有职官系统而告结束”。

——《唐六典》

史料9:

>张国刚曾指出,“差遣期间本官是不管本司事的,原来的职事官只是表示该官受差遣时的身份地位而已,这样一来,就埋下了职事官阶官化的种子”。

——张国刚《道宣集》

史料10:

>《通典》描述,安禄山和哥舒翰督军期间,士卒普遍被授予中低级武职事官衔,甚至超出典制规定的员额,形成了一种官称的泛滥状况。

——《通典》

当前观点（有辨别难度的错误观点）

唐代武职事官体系一直保持名实一致，其名号和职责始终匹配，并未出现职位虚衔化的问题。

反驳当前观点的历史推理过程

田归道在担任数个实职的同时还兼任其他职责（史料1）+安禄山时期士卒被授中低级武职事官衔，但职位泛滥严重，并未实际履行相应职责（史料3）->唐代并未始终保持武职事官的名号和职责一致，特别是在特定时期出现了职位虚衔化现象（推论1）

推论1+张国刚指出差遣期间本官不管本司事，原职事官仅为身份地位象征（史料7）->唐代武职事官的职事官虚衔化问题确实存在，名号与职责不匹配的情况尤为明显（反驳错误观点）

#####

Output:

分析

- 观点：当前观点是错误的，但判断难度较低，包含“一直”、“始终”、“并未”等降低题目难度的绝对性用语，需要修改。

- 历史推理过程：历史推理过程存在一定冗余，史料1、史料3和史料7的意义相近，需要重新组织历史推理过程使其符合要求。

修改后的观点（必须仍然是有辨别难度的错误观点）

嘉靖时期的明世宗是明朝的“中兴”之主，张居正认为他的革新之法值得后世效法。

修改后的历史推理过程

田归道在担任数个实职的同时还兼任其他职责（史料1）+安禄山时期士卒被授中低级武职事官衔，但职位泛滥严重，并未实际履行相应职责（史料10）+张国刚指出差遣期间本官不管本司事，原职事官仅为身份地位象征（史料9）->唐代存在武职事官的职事官虚衔化问题，名号与职责不匹配的情况尤为明显（反驳错误观点）

#####

Example 3:

#####

Input:

史料列表

史料1:

>根据调查，景德镇传统民窑的生产技术、产品质量和生产成本在一定程度上受到了御窑的影响与带动。

——王世襄《清代匠作则例汇编》

史料2:

>影响因素的敏感度分析中，价格和成本对景德镇瓷业实际产值的敏感度最高，特别在晚清时期。清朝末期所有要素的波动较大，价格的波动尤为显著。

——《景德镇陶瓷概况》

史料3:

>清后期瓷器包装物料价格大幅上涨，这解释了清代瓷业产值高峰出现在光绪朝的原因。

——《清宫瓷器档案全集》

史料4:

>咸丰至同治初年是景德镇瓷器烧造数量的低谷阶段，但光绪朝后，生产数量相比前期大幅度减少，且清末景德镇单件瓷器的生产成本大幅提高，光绪时期单件瓷器的生产成本约为雍乾时期的6倍左右。

——《清宫瓷器档案全集》

史料5:

>瓷器生产受到当地的天气影响，如果天气寒冷，泥土干冻，就无法进行瓷器生产，开工时间会延后。清代御窑生产过程中有一标准环节名为“大开工祭祀”，即开工之前的祭祀活动。

——《清宫瓷器档案全集》

史料6:

>雍正十三年年的供词中明确提到当时的御窑生产已经转为厂内制作坯胎民窑进行烧造，说明御窑与民窑的生产是有紧密联系的。

——舒善《供词》

史料7:

>清宫瓷器档案所载从雍正四年至宣统三年的御窑瓷器详细记录, 这些记录包括了各朝各年的御窑生产情况、成本与燃料费用等详细数据。

——铁源、李国荣《清宫瓷器档案全集》

史料8:

>高岭土对瓷器的生产至关重要, 清代开采和使用的高岭土主要来源地有浮梁东埠高岭村、黄坛大洲村和九江星子; 其中东埠土被替代, 增加了开采成本和运输成本, 影响了清代瓷器的产值。

——《景德镇陶瓷史》

史料9:

>道光二十九年, 景德镇的开工祭祀延迟了八天, 冬季平均气温每降低1摄氏度, 大开工祭祀的时间就会晚8.84天。

——《江西通志》

当前观点 (有辨别难度的错误观点)

光绪朝之后景德镇瓷器生产数量的减少主要是由于御窑与民窑的密切联系减弱, 导致民窑生产技术和产品质量下降, 进而影响了整体生产数量, 而不是由于生产成本的提升。

反驳当前观点的历史推理过程

御窑与民窑的生产联系在雍正时期之后非常密切 (史料6) + 景德镇民窑的生产技术、产品质量和生产成本受御窑的影响 (史料1) -> 光绪朝之后御窑与民窑的密切联系若减弱, 可能会导致民窑生产技术和产品质量的下降 (推论1)

光绪朝之后景德镇瓷器生产数量减少主要是由于生产成本的大幅上升 (史料4) + 影响因素中的价格和成本对瓷业产值的波动显著, 尤其在晚清时期 (史料2) -> 光绪朝之后景德镇瓷器生产数量减少主要是由于生产成本大幅上升, 而非因御窑与民窑联系减弱 (推论2)

推论1 + 推论2 -> 光绪朝之后景德镇瓷器生产数量减少并不是由于御窑与民窑的联系减弱导致生产技术和质量下降, 而是因生产成本上升 (反驳错误观点)

#####

Output:

分析

- 观点: 当前观点是错误的, 但判断难度较低, 需要修改, 去掉“而不是由于生产成本的提升”后当前观点仍然是错误的, 且辨别难度更高。

- 历史推理过程: 根据修改后的观点, 历史推理过程中的相关部分进行相应修改, 使符合人类逻辑、历史常识和推理约束, 能够基于该历史推理过程可以反驳修改后的当前错误观点。

修改后的观点 (必须仍然是有辨别难度的错误观点)

光绪朝之后景德镇瓷器生产数量的减少主要是由于御窑与民窑的密切联系减弱, 导致民窑生产技术和产品质量下降, 进而影响了整体生产数量。

修改后的历史推理过程

御窑与民窑的生产联系在雍正时期之后非常密切 (史料6) + 景德镇民窑的生产技术、产品质量和生产成本受御窑的影响 (史料1) -> 光绪朝之后御窑与民窑的密切联系若减弱, 可能会导致民窑生产技术和产品质量的下降 (推论1)

光绪朝之后景德镇瓷器生产数量减少主要是由于生产成本的大幅上升 (史料4) + 影响因素中的价格和成本对瓷业产值的波动显著, 尤其在晚清时期 (史料2) -> 光绪朝之后景德镇瓷器生产数量减少主要是由于生产成本大幅上升 (推论2)

推论1 + 推论2 -> 光绪朝之后景德镇瓷器生产数量的减少可能与御窑与民窑的联系减弱导致生产技术和质量下降有关, 但这不是主要原因, 主要原因是生产成本的上升 (反驳错误观点)

#####

-Real Data-

#####

记住, 每一个推理符“->”“左侧的节点**至少有2个节点**, 每一行仅有一个推理符”->“禁止一个节点解释多个史料!!! 这对我的工作非常重要。

A.2.2 HR-TUNING DATA GENERATION

Source Interpretation

-Goal-

历史是一门注重逻辑推理和严密论证的实证性的人文社会学科。现在，给定特定史料，您要尝试解读其含义。

-Input Format-

一个一手或二手的史料

-Output Format-

史料含义解读

#####

-Real Data-

#####

Source Deduction

-Goal-

历史是一门注重逻辑推理和严密论证的实证性的人文社会学科。现在，给定多个特定史料，尝试在这些史料的基础上进行联合推理。

-Input Format-

多个一手或二手的史料

-Output Format-

史料1的解释（史料1）+ 史料2的解释（史料2）+ ... ->推论内容（推论）

#####

-Real Data-

#####

A.3 CORRECTION INSTRUCTIONS

A.3.1 CORRECTION FORMAT

Directly modify the option content or historical reasoning process in the “question.txt” file. Mark the modified option content or historical reasoning process with an asterisk (*) at the end of the line; no additional reasons for modification are needed.

A.3.2 CORRECTION LOGIC

- a) Only modify the option content or historical reasoning process; do not change the answers (i.e., do not change ABC to AC).
- b) The true option is mostly the viewpoint of the paper’s author and is generally correct. However, if there is insufficient historical evidence to infer the option, it can be slightly modified or narrowed in meaning to make it inferable based on the existing corpus.
- c) false options may sometimes be too simple, containing absolute expressions (e.g., “completely”, “always”) or many negative expressions (e.g., “did not”). Try to remove these expressions or soften the tone to make the question more challenging.
- d) The historical reasoning process may contain some leaps or require additional knowledge, but if it still aligns with logic or historical common sense, it is considered acceptable.

A.4 MORE REASONING SAMPLES FROM THE SYSTEM

Here are some samples we generated from LLMs. Judging from the following samples, our generated historical questions are good in difficulty and reliability. They require analysis of multiple sources and

consideration of factors. Answers are well-reasoned and evidenced, following the detailed historical reasoning processes. This indicates our data generation process is effective in creating high-quality questions for LLMs improvement.

A.4.1 SAMPLE #1

不定项选择题

阅读下列史料，判断下列选项哪些最有可能成立：

史料1：

仲几父簋记载器主受命“使于诸侯、诸监”，可见地方政区除了有军政主官“侯”外，还设有负责督察的官员“监”。

——《集成》

史料2：

大国三卿，皆命于天子……次国三卿，二卿命于天子，一卿命于其君。

——《礼记·王制》

史料3：

从前文已称“文”“武”之谥看，也应指成王。而此后历代周王虽然也有分封之举，但均零星分散，诸侯功能和取事与周初已大有不同。

——《系年》

史料4：

文献中有成王时周、召“分陕而治”《春秋公羊传》隐公五年云：“自陕而东者，周公主之；自陕而西者，召公主之。”

——《春秋公羊传》

史料5：

《四告》中所“建侯”即包括“出分子”，亦与之同谓。从文献看，周初分封的诸侯大部分是与周王血缘极近的姬姓贵族。一些外戚也被分封为诸侯，虽不乏齐、申等地处东、西二土的“周之厚屏”，但总体力量较姬姓诸侯仍有不小差距。

——《四告》

史料6：

监于夏商之既败，丕则亡遗后，至于万亿年。

——《清华简》

史料7：

清华简存有成王册封吕丁于许的命书《封许之命》，篇中有“今朕永念乃勋，命汝侯于许”一句，是成王自述念及吕丁在文、武两朝勤劳王事，命他到许地做“侯”。

——《清华简》

史料8：

地方封国军队由诸侯负责训练和管理，但仍是周天子的武装力量。无论是诸侯还是公卿，都只是替周王统领部队。

——《国语·鲁语下》

史料9：

清华简《祭公之顾命》回顾分封之事：我亦上下辟于文武之受命，广戡方邦，丕惟周之旁，丕惟后稷之受命是永厚。惟我后嗣，方建宗子，丕惟周之厚屏。

——《清华简》

史料10：

西周晚期的师簋记载周王室打击淮夷的一次军事行动，器主师受王命率领齐、纪、莱等山东诸侯军事力量联合作战。

——《集成》

史料11：

宜侯簋记载，周王命为宜侯时，在省闾“伐商图”“东国图”等进行“赐土”与“赐人”。

——《宜侯簋》

史料12：

此铭前半段为：唯正月初吉丁亥，王格于穆宫，桓伯佑左右即位。王命左右曰：“更乃祖考作家司工于蔡……”

——《左右簋》

史料13：

地方封国部队从某种意义说更像是中央驻军。周王不但可以直接调动、指挥诸侯部队，还可以派员在驻地执掌。

——引簋，《铭图》

选项：

- A. 中央通过任命和监督地方诸侯的卿大夫，直接掌控地方的政治和行政事务，确保了对地方封国的有效控制。
- B. 西周中央权力的高度集中体现在其能够调动和控制地方封国的军事力量，确保对地方的军事掌控。
- C. 周初分封制的实施除了吸收前朝教训，还体现了浓厚的血缘和宗法观念，使得周初分封带有极强的政治与军事目的。
- D. 周初分封制度设计通过宗亲和外戚的委任，增强了对地方的控制，使地方封国形成了类似于中央行政机构的组织结构。

正确答案：

ABCD

历史推理过程：

1. 选项A：

大国三卿由天子直接任命，次国由天子任命二卿，一卿由其君任命（史料2）+地方的“侯”由周王直接任命（史料7）->周王对地方封国的人事任免权深入到诸侯宰辅层级（推论1）

推论1+地方政区除了诸侯外，还有负责监督的监官（史料1）+周王能够直接调动和指挥诸侯部队（史料13）

->中央通过任命和监督地方诸侯的卿大夫，直接掌控地方的政治和行政事务，确保了对地方封国的有效控制（选项A正确）

2. 选项B：

地方封国军队名义上由诸侯管理，但实际上替周王统领（史料8）+中央可以直接调动和指挥地方封国的军事力量（史料10）->中央对地方军队的控制确保了对地方的军事掌控（推论1）

中央派员驻守和管理地方军事力量（史料13）+周王通过分陕而治的方式加强对地方的控制（史料4）->中央对地方封国的军事力量进一步控制的实例（推论2）

推论1+推论2->西周中央权力的高度集中体现在其能够调动和控制地方封国的军事力量，确保对地方的军事掌控（选项B正确）

3. 选项C：

夏商最终被外服邦伯取代，引起周人警醒（史料6）+周初大规模分封由周公于成王世推动（史料3）->周初实施分封制吸收前朝覆亡教训（推论1）

周初分封的诸侯多为血缘极近的姬姓贵族，并辅以外戚（史料5）+历代周王分封活动零星散落，职事变化（史料3）->周初分封显示出浓厚的血缘和宗法观念（推论2）

推论1+推论2->周初分封制的实施除了吸收前朝教训，还体现了浓厚的血缘和宗法观念，使得周初分封带有极强的政治与军事目的（选项C正确）

4. 选项D：

分封诸侯多为血缘极近的姬姓贵族及少量外戚（史料5）+周王在分封时既赐土又赐人以确保新封区域的稳定管理（史料11）->分封制度增强了对地方的控制（推论1）

周王室通过分封制形成对地方封国的控制（史料10）+宗亲和外戚在分封过程中受命管理地方，形成了周之厚屏（史料9）->分封制度使地方封国形成类似于中央行政机构的组织结构（推论2）

推论1+推论2->周初分封制度设计通过宗亲和外戚的委任，增强了对地方的控制，使地方封国形成了类似于中央行政机构的组织结构（选项D正确）

A.4.2 SAMPLE #2

不定项选择题

阅读下列史料，判断下列选项哪些最有可能成立：

史料1

宋太宗即位后，为统一全国，除灭北汉和伐幽州外，还在太平兴国七年(982)抓住机会取缔了夏州党项政权。这一举动激起了以李继迁为首的党项部族反抗，特别是李继迁将战略目标指向灵州后，宋军在西北战场开始陷入被动。

——《中国边疆史地研究》

史料2

学者指出，李沆当政时期，曾极力摈抑曾致尧、梅询，曾致尧等人与张齐贤均主张坚守灵州，而曾受李沆提携的杨亿，与李沆同样主张放弃灵州。

——《崇高基行之于介休而效，马丕瑤行之于解州而益效》

史料3

宋代弭兵论兴起并成为主流思潮的时代背景，是宋太宗对辽战争的转折性失利。雍熙北伐之后，宋朝在今河北、山西一线面临巨大压力，战争消耗巨大，边地百姓饱受战争之苦。于是太宗朝后期，“姑务羸縻，以缓争战”的弭兵之策顺势而起。

——《晚清山西土地清丈》

史料4

学界多认为这一方针从属中国边疆史地研究年第一期于宋朝“守内虚外”“崇文抑武”的政策，即文官政治的产物。不过，作为宋初弭兵论研究的奠基之作，陈芳明的论文早就注意到一个现象，张齐贤虽然在契丹事务上倡导弭兵论，但在对党项事务中，他却力主坚守灵州，采取强硬政策，李沆则与之针锋相对，主张放弃灵州。

——陈芳明《张齐贤与李沆之争》

史料5

咸平四年九月清远军失陷后，知宁州石熙政上言，矛头直指主弃派，“盖朝廷素不留意”，知枢密院事周莹斥为“岂得如此悖慢”。十月张齐贤自陕西使还，系统阐述自己对西北战局的意见，主张以重爵联结凉州蕃部，共同对付党项。

——《宋代弭兵论研究》

史料6

特别是李继迁将战略目标指向灵州后，宋军在西北战场开始陷入被动。尤其是至道二年(996)，五路大军西征党项未竟全功，以及宋真宗继位后朝廷政策全面转向，都使宋朝逐渐对党项采取绥靖政策。

——《王夫之批评宋真宗》

史料7

张齐贤于咸平元年十月与李沆同时拜相，至三年十一月罢相，其间与李沆、王旦、寇准、张咏等人多有角力。张在当年十一月因醉酒失仪罢相，极可能是因为时任言官的李沆恩主、王旦岳父赵昌言的弹劾。

——《宋史》

史料8

咸平二年(999)秋，李继迁集中兵力东向进攻麟州，重创劲敌府州折氏。次年九月，宋军辎重部队又在瀚海遭到党项围歼，知灵州李守恩与陕漕陈纬等边防要员阵亡，此事震动朝野。

——《宋史》

史料9

十月丁未(九日)，张齐贤等人回朝上奏，提出自己的意见。他指出现在陕西最大的问题是兵力不足，但河北防线压力太大，无法支援，最好的办法是“激励自来与继迁有仇蕃部，招诱远处大族首领”。

——《宋史》

史料10

张齐贤在咸平五年灵州围解之时，张齐贤提出通过重赏引导凉州蕃部共同对抗党项，但未被采纳。

——宋代史料

选项：

- A. 西北弭兵论是宋代士人群体内部派系斗争的产物，并不仅仅是文官政治的体现。
- B. 灵州的守弃之争是朝廷内主要派系（主守派和主弃派）之间政治博弈的具体体现。
- C. 李沆与张齐贤之间的派系斗争延误了灵州的援救工作，导致西北危局的恶化。
- D. 主弃派与主守派的异议不仅是政见不同，更反映了当时的政治派系斗争。

正确答案：ABCD

历史推理过程：

1. 选项A：

弭兵论作为宋太宗后期的政策有其时代背景（史料3）+ 学界多认为这一方针从属于宋朝“守内虚外”“崇文抑武”的政策，即文官政治的产物。（史料4）->传统观点认为弭兵论属于文官政治的产物（推论1）
李沆与张齐贤在具体党项事务中存在严重分歧，并且存在派系斗争（史料2）+ 张齐贤在党项事务中主张坚守灵州，而李沆主张放弃灵州（史料4）->西北弭兵论实际上是宋代士人群体内部派系斗争的产物（推论2）
推论1 + 推论2 ->西北弭兵论是宋代士人群体内部派系斗争的产物，并不仅仅是文官政治的体现。（选项A正确）

2. 选项B：

李继迁将战略目标指向灵州，宋军在西北陷入被动（史料1）+ 宋军辎重部队在瀚海遭党项围歼，知灵州李守恩等边防要员阵亡（史料8）->灵州成为双方争夺的重要战略要地（推论1）
张齐贤主张联合凉州蕃部援助灵州，并因此与主弃派产生矛盾（史料5）+ 李沆和张齐贤在是否守灵州的问题上有重大分歧（史料2）->朝廷内部存在主守派和主弃派的分歧（推论2）
清远军失陷后，知枢密院事周莹斥责主弃派（史料5）+ 党项部袭击灵州引发朝野震动（史料8）->灵州的守弃问题引发朝廷主要派系之间的政治博弈（推论3）
推论1 + 推论2 + 推论3 ->灵州的守弃之争是朝廷内主要派系（主守派和主弃派）之间政治博弈的具体体现。（选项B正确）

3. 选项C：

李继迁将战略目标指向灵州后，宋军被动（史料6）+ 张齐贤提出通过联合外部势力援救灵州（史料9）->张齐贤主张援助灵州的策略以解决危局（推论1）
李沆派系的弹劾最终导致了张齐贤被罢相，两人之间存在严重的派系斗争（史料7）+ 李沆在对党项事务上的不同立场，主张放弃灵州（史料4）->派系斗争严重影响了援救灵州工作的展开（推论2）
推论1 + 推论2 ->李沆与张齐贤之间的派系斗争延误了灵州的援救工作，导致西北危局的恶化。（选项C正确）

4. 选项D：

张齐贤主张坚守灵州（史料10）+ 杨亿主张放弃灵州（史料2）->主弃派与主守派在灵州问题上存在政见分歧（推论1）
张齐贤提出通过重赏引导凉州蕃部共同对抗党项，但未被采纳（史料10）+ 李沆大力摈抑主守派成员（史料2）->主弃派与主守派的分歧反映了深刻的政治派系斗争（推论2）
推论1 + 推论2 ->主弃派与主守派的异议不仅是政见不同，更反映了当时的政治派系斗争（选项D正确）

A.4.3 SAMPLE #3

不定项选择题

阅读下列史料，判断下列选项哪些最有可能反映了嘉靖时期边商与内商分工的原因。

史料1

庞尚鹏在隆庆二年（1568）清理盐法报告中提到：

查得国初原无边商、内商名色。自边商难于守支，故卖引于内商。内商难于报中，故买引于边商。一专报中，一专守支。

——《清理盐法疏》

史料2

《明史·食货志》记载：

成化十九年颇减存积之数，常股七分，而存积三分。然商人乐有见盐，报中存积者争至，遂仍增至六分。淮、浙盐犹不能给，乃配支长芦、山东以给之。一人兼支数处，道远不及亲赴，边商辄贸引于近地富人。自是有边商、内商之分。

——《明史·食货志》

史料3

董光裕在万历六年（1578）指出：

成化年间，每引派纳折银三钱五分，递加至四钱二分。当此之时，每一边商引盐派定一运司，未有搭派别处之例，故商人亲中亲支，随支随掣，淮南岁掣八单，淮北四单，引盐流通，绝无阻格，亦无内商、边商之别。

——董光裕《议禁掣河盐》

史料4

《明实录》记载：

隆庆元年（1567）巡盐御史苏朝宗奏曰：“国初边商亲自支盐，至仪、淮二所掣卖。其后困于余盐，将河盐堆置淮扬，存积渐多，不暇守候，乃分拨引目，鬻之居民。故内商坐致富饶，而边商奔走益困。”

——《明实录》

史料5

《明史》记载：

嘉靖三十二年，工本盐加增边商负担，商人不堪其苦，纷纷将盐引售于内商，遂使内商日益富有。

——《明史》

选项：

- A. 嘉靖时期盐政改革的推行使得边商与内商的分工制度化，强化了内商的地位。
- B. 成化年间的配支政策是边商与内商分化的主要原因，并在嘉靖时期通过工本盐制度得以巩固。
- C. 边商与内商的分化在成化年间就已完全实现，嘉靖时期的盐政改革仅是对此分化的延续和确认。
- D. 工本盐制度的推行虽然导致了边商与内商的分工，但真正制度化分工的关键还是在于边商无法支盐的困境。

正确答案：

AD

历史推理过程：

1. 选项A：

嘉靖三十二年工本盐加增边商的负担，使边商将盐引售于内商（史料5）+ 内商通过这种方式日益富有，最终形成了制度化的分工（史料4）->嘉靖时期工本盐等盐政改革的推行使得边商与内商的分工制度化，强化了内商的地位（选项A正确）

2. 选项B：

成化年间的配支政策导致了一定程度的边商与内商的分化（史料2）+ 成化年间商人仍能亲自报中支盐，说明分化并未完全实现（史料3）->无法确认成化年间的配支政策是边商与内商分化的主要原因（选项B错误）

3. 选项C：

成化年间未实现内商与边商的分化（史料3）+ 嘉靖时期（隆庆元年前）完成了内商与边商的分化（史料4）->分化的制度化可能是在嘉靖时期完成的，成化年间并未完全实现这种分化（选项C错误）

4. 选项D：

边商难于支盐导致的分工（史料1），工本盐加重了边商的负担，使得他们无法支盐，只能将盐引卖给内商，从而推动了分工的制度化（史料5）->工本盐制度导致了边商与内商的分工，但分工的关键在于边商无法支盐的困境（选项D正确）

ACKNOWLEDGEMENTS & RESEARCH PROCESS

Artificial Intelligence (AI) has been quite a hot topic for the last several years, especially with the surge of Large Language Models (LLMs) such as ChatGPT. Additionally, history and the humanities in general have been subjects where I've been passionate for a long while. Recently, the development of said AI technology has opened up several fields that fuse together AI and machine learning with different disciplines and industries, such as AI for healthcare, AI for Engineering, AI for Science, etc. One field that has yet to be explored adequately is AI for History, a denomination of the AI for Social Sciences Branch of AI fusion disciplines.

Due to these passions, with my parents' help, I sought out Dr. Zhi-Qi Cheng from CMU and Prof. Jun Xiao from Zhejiang University for guidance on researching this novel field. Prof. Xiao assisted with the ideation of the project and guidance in topic exploration. Dr. Cheng guided me throughout the whole project, providing consistent, timely, and valuable feedback on my process in this endeavor. There were no other helpers in the process of this project, except in the seed data acquisition and in the human verification steps for question generation, which are both non-technical parts. With the help of Dr. Cheng, we collected around 1800 historical research papers published in recent academia journals on the topic of Chinese history. The human historians verified around 300 problems that were generated by AI. The paper was solely written by me (as this was a solo project without any teammates), however, there were several rounds of commentary feedback from Dr. Cheng and Prof. Xiao before the final version.

The original topic of this project was to create a Historical LLM Chat model in order to synthesize arguments, have scholarly conversations, and create a paper/argument summary based off of the conversation with a human user. However, this idea was found to be too big, therefore, I, with the assistance of the two professors, narrowed it down to one of the most difficult aspects of historical research – reasoning. This was, however, still too large of a scope, so we narrowed it further to evidence reasoning, which is a more distilled version of the aforementioned difficulty. Upon further exploring, we've found that there is no dataset or test data for a baseline to measure our progress. Without a baseline, it would be extremely difficult to find out whether there can be any improvement or progress in this chosen field.

The project had multiple steps where Dr. Cheng was instructing. In the first step, seed data collection, he utilized resources from Zhejiang University to download high-quality scholarly research papers regarding Chinese history from 2020-2024. Using these downloaded papers, I was able to begin to engineer a prompt that would use this seed data to generate questions. The prompt engineering was solely done by me, including writing, testing, analysis, and editing. Once this prompt engineering process was complete and around 1800 questions were generated, several human experts verified around 300 problems - as mentioned above as around 90% accurate and high-quality. After this process, I've utilized these high quality questions (excluding the human-verified questions, which will be used to do final testing) to conduct so-called HR-tuning (supervised fine tuning for historical reasoning) on different LLM models to test their ability on solve problems, i.e., the ability of historical reasoning with evidence. During this process, I've encountered many problems and bugs both in the data generation process and the tuning process, and Dr. Cheng assisted in some problem resolving and code debugging, although all of the final execution and tuning was done by me. Nonetheless, in the end, after several tedious rounds of debugging and improving, I was able to achieve success in having some models score as high as 65.55% accuracy in answering the generated and verified questions. This is significantly high than an AI models without tuning.

After this, during the period of paper writing, the outline and creation of the paper was done solely by me. Several rounds of feedback occurred between Dr. Cheng, Prof. Xiao and I to enhance the paper and ensure all findings were objective, accurate, and fairly presented in a clear and organized manner.

I'd like to thank both Dr. Cheng and Prof. Xiao for their helpful and insightful comments on my topic selection and paper writing, in which they've significantly helped to improve my understanding of scientific research practices, skillful communication, and accurate representations of data. I'd also like to thank Dr. Cheng for his directing and instructing throughout the entire research process that helped me to push forward despite significant adversity. I'd like to thank the digital age for allowing me to access so much high-quality information and open-source AI models at my fingertips to empower myself to conduct this research. Finally, I'd like to acknowledge my parents unwavering support for my research. Without all this support, this research wouldn't have been possible.

BIO OF ADVISORS

DR. ZHI-QI CHENG

Dr. Zhi-Qi Cheng is a Project Scientist at the Language Technologies Institute (LTI), which is part of the School of Computer Science at Carnegie Mellon University (CMU). He earned his B.S. in Computer Science from Southwest Jiaotong University in 2014 and completed his Ph.D. in 2019. During his doctoral studies, he was a visiting Ph.D. student at the City University of Hong Kong (2016-2017) and later at CMU (2017-2019). He also gained valuable experience through internships at Alibaba DAMO Academy (2016), Google Brain (2018), and Microsoft Research (2019). From 2019 to 2022, he served as a postdoctoral research associate at CMU. During his time at CMU, he made significant contributions to several key projects, including DARPA's AIDA, KAIROS, IARPA's DIVA, and NIST's PSIAP. Notably, on the DARPA KAIROS project, he was solely responsible for building the system for the entire CMU team. He has published numerous papers in top conferences such as CVPR, ICCV, NeurIPS, IJCAI, and ACM MM. His research has had a substantial impact, such as being utilized in the Washington Post's coverage of the Capitol riots, for which he was awarded the Pulitzer Prize for Public Service. Additionally, he has been honored with the Intel Ph.D. Fellowship and IBM Outstanding Student Scholarship. More information can be found at <https://zhiqic.github.io/homepage/>.

肖俊教授

浙大计算机学院人工智能研究所教授/博士生导师/副所长，兼任浙江大学数字法治研究院副院长，中国人工智能产业发展联盟开源开放推进组秘书长，教育部“长江学者奖励计划”特聘教授，入选2022年全球前2%顶尖科学家（人工智能与图像处理领域）及爱思唯尔2023、2022年度中国高被引学者，国家科技创新2030-新一代人工智能重大项目“新一代人工智能科教创新开放平台”负责人/首席科学家，2019年入选浙江省“万人计划”科技创新领军人才，获浙江省科技进步一等奖（2021）及浙江省自然科学奖一等奖（2015），2018年获得浙江省自然科学基金杰出青年基金资助并入选浙江省151人才工程第二层次培养人员，2015年入选浙江大学求是青年学者。担任IEEE TPAMI、IEEE TMM、IEEE TCSVT、Information Science、Signal Processing、Neurocomputing等国际期刊审稿人，以及ACM MM、AAAI、CVPR、ICCV、ICML、ICLR、NeurIPS等高水平国际学术会议AC/SPC/PC或审稿人；获得多媒体领域全球顶级学术会议ACM MultiMedia 2023唯一最佳论文奖；先后在IEEE TKDE、IEEE TMM、IEEE TVCG、IEEE TCB、CVPR、ICCV、ACM MM、IJCAI、AAAI、SIGIR、Information Science、Neurocomputing、Pattern Recognition、Signal Processing等重要国际学术期刊和顶级学术会议上发表（录用）论文100余篇；获得国家发明专利授权50余项，其中多项进行了转让或授权实施；拥有软件著作权30余项。更详细的信息请参阅：<https://person.zju.edu.cn/junx>