Large-Scale Corpus Construction and Retrieval-Augmented Generation for Ancient Chinese Poetry: New Method and Data Insights

Yang Liu[1,3], **Lan Lan**[1], **Jiahuan Cao**[1,3], **Hiuyi Cheng**[1,3], **Kai Ding**[2,3], **Lianwen Jin**[1,3,4*]
[1]South China University of Technology
[2]INTSIG Information Co., Ltd
[3]INTSIG-SCUT Joint Lab on Document Analysis and Recognition
[4]SCUT-Zhuhai Institute of Modern Industrial Innovation
ly10061105@gmail.com, eelanlan@mail.scut.edu.cn, jiahuanc@foxmail.com
eechenghiuyi1@mail.scut.edu.cn, danny_ding@intsig.net, eelwjin@scut.edu.cn

Abstract

Ancient Chinese Poetry (ACP), a critical aspect of Chinese cultural heritage, presents unique challenges for Large Language Models (LLMs). One of the most pressing challenges is the significant hallucination issues faced by LLMs due to data scarcity and limited ability of general LLMs when dealing with ACP. To address these challenges, this paper constructs the ACP-Corpus, which encompasses 1.1 million ancient poems and 990K related texts, designed to enhance the training and performance of LLMs. Alongside this, we develop the ACP-QA dataset, comprising over 12 million question-answer pairs across 24 task categories, and the ACP-Eval dataset for rigorous evaluation purposes, containing 7,050 entries. Building on this resources, we propose the ACP-RAG framework, a specialized Retrieval-Augmented Generation (RAG) approach that significantly improves the performance of LLMs in the domain of ancient poetry from 49.2% to 89.0%. The ACP-RAG contains five modules of semantic coarse-grained retrieval, semantic fine-grained retrieval, keyword retrieval, keyword matching, and context filtering. Experiments show that ACP-RAG achieves a promising response accuracy of 89.0%, surpassing existing LLMs by a remarkable margin. We believe this work not only advances the capabilities of LLMs in processing ancient Chinese poetry but also contributes to the preservation and innovative development within this rich literary tradition. The datasets and code are available at https://github.com/SCUT-DLVCLab/ACP-RAG.

1 Introduction

As a treasure of Chinese culture, ancient Chinese poetry embodies a rich tapestry of history, culture, and emotion, representing one of the significant cultural heritages of the Chinese nation. Prior to the emergence of Large Language Models (LLMs), the integration of ancient poetry and artificial intelligence primarily focused on sentiment classification and poetry generation. Notable studies in this area include sentiment analysis by Chen et al. (2019) and Sheng and Uthus (2020), and poetry generation by He et al. (2012) and Yi et al. (2020). Despite these efforts, the methods often fall short in capturing the nuanced cultural connotations and artistic essence of ancient poetry.

The advent of LLMs has unlocked new potentials for promoting ancient poetry. Leveraging their advanced Natural Language Processing (NLP) capabilities, LLMs offer significant promise in this specialized field. However, challenges such as data scarcity specific to ancient poetry, a lack of specialized models, and prevalent hallucination issues in current LLMs (Liu et al., 2024b) continue to limit their effectiveness. For instance, the average score of eight LLMs on the ACLUE benchmark is only 32.6 (Zhang and Li, 2023).

Currently, many existing studies mitigate hallucination issues through Retrieval-Augmented Generation (RAG) (Gao et al., 2023). Accordingly, we aim to develop a RAG system tailored for the field of ancient poetry, aiming to enhance LLM capabilities and address data deficiencies in this domain. Unlike most RAG methods designed for general documents, our study focuses on this unique vertical domain of ancient Chinese poetry.

The reason we choose RAG over Supervised Fine-Tuning (SFT) is that the RAG system has relatively lower hardware requirements during training (only two 3090 GPUs), while SFT requires large-scale computational resources for full fine-tuning, placing higher demands on hardware.

To this end, we have developed the **ACP-Corpus**, which contains 1,124,024 poems and 44,347 authors, covering 26 dynasties and 722 themes, and enriched with 990,801 related texts including literary appreciations, translations, and id-

* Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 779–817
April 29 - May 4, 2025

Corpus	# Poems	# Authors	# Dynasties	# Themes	Other Knowledge Texts						
					ET	PA	VT	WE	ID	LK	PI
THU-CCPC (Guo et al., 2019)	127,682	8,111	9	-	✘	✘	✘	✘	✘	✘	✘
Chinese-Poetry (Chinese-Poetry, 2017)	396,170	18,789	8	-	✘	✔	✘	✘	✘	✘	✘
Poetry (Werneror, 2017)	793,321	28,387	16	-	✘	✘	✘	✘	✘	✘	✘
Chinese-poetry-and-prose (VMIJUNV, 2022)	835,564	38,418	16	-	✘	✔	✔	✔	✘	✘	✔
ACP-Corpus (Ours)	**1,124,024**	**44,437**	**26**	**722**	✔	✔	✔	✔	✔	✔	✔

Table 1: Comparison of existing pre-training corpora. "ET" indicates English Translation, "PA" indicates Poem Appreciation, "VT" indicates Vernacular Translation, "WE" indicates Word Explanation, "ID" indicates Idioms, "LK" indicates Literary Knowledge, and "PI" indicates Poet Introduction.

Dataset	Domain	License	Scale	# Tasks	# IC	Method		
						HG	CI	MC
Firefly (Yang, 2023)	General	-	100,845	2	2	✔	✔	✘
COIG-CQIA (Bai et al., 2024)	General	-	391	3	2	✔	✔	✘
ACCN-INS (Cao et al., 2024b)	Classical Chinese	CC BY-NC-SA-4.0	7,767	12	4	✔	✔	✔
ACP-QA (Ours)	Ancient Chinese Poetry	CC BY-NC-SA-4.0	**12,571,851**	**24**	**5**	✔	✔	✔

Table 2: Comparison of existing Q&A datasets. "IC" indicates Instruction Categories, "HG" indicates Human Generated, "CI" indicates Collection and Improvement of existing datasets, and "MC" indicates Model Constructed.

ioms. Building on this corpus, we have categorized tasks into 24 distinct categories and generated a substantial dataset of 12,571,851 Question-Answer (Q&A) pairs, termed **ACP-QA**. This dataset can serve not only as a database for RAG but also for instruction fine-tuning. Additionally, we create a evaluation dataset of 7,050 entries in the field of ancient poetry, named **ACP-Eval**.

Building on this foundation, we develop the **ACP-RAG** framework for the field of ancient poetry, which includes five modules: semantic coarse-grained retrieval, semantic fine-grained retrieval, keyword retrieval, keyword matching, and context filtering. Comparative experiments demonstrate that ACP-RAG improves the correctness of LLMs' answers from 49.2% to 89.0%, confirming its effectiveness. To comprehensively evaluate the RAG system, we establish six evaluation metrics for the experiments in this paper. Additionally, we specifically fine-tune a scoring model to evaluate these metrics.

The main contributions of this paper are as follows:

- We build a comprehensive ancient poetry corpus, ACP-Corpus, the Q&A dataset ACP-QA, and the evaluation dataset ACP-Eval.
- We propose a novel Retrieval-Augmented Generation framework, ACP-RAG, tailored for ancient Chinese poetry, incorporating advanced retrieval and matching techniques to significantly enhance LLM performance.
- We introduce six metrics to conduct a more comprehensive evaluation of the RAG system. Additionally, we specifically fine-tune a scoring model for the evaluation.

2 Related Work

2.1 Ancient Chinese Poetry Data

The ancient Chinese poetry data is mainly divided into three categories: Pre-training Corpora, Instruction Fine-Tuning Datasets and Evaluation Datasets.

(1) Pre-training Corpora. The ancient poetry corpora include four main datasets: Chinese-poetry (Chinese-Poetry, 2017), Poetry (Werneror, 2017), Chinese-poetry-and-prose (VMIJUNV, 2022), and THU-CCPC (Guo et al., 2019).

(2) Instruction Fine-Tuning Datasets. The ACCN-INS dataset (Cao et al., 2024b) focuses on classical literature and includes 7,767 instructions related to ancient poetry. The Firefly dataset (Yang, 2023) includes 69,950 instructions concerning the generation of ancient poetry.

(3) Evaluation Datasets. The WenMind (Cao et al., 2024a), WYWEB (Zhou et al., 2023), and ACLUE datasets (Zhang and Li, 2023) serve as benchmarks for evaluating classical Chinese, comprising tasks such as poetry appreciation. Furthermore, the CCPM dataset (Li et al., 2021) focuses on translation task, while the THU-FSPC dataset (Chen et al., 2019) concentrates on sentiment classification task.

2.2 Retrieval-Augmented Generation

RAG utilizes external knowledge bases to provide contextual information for LLMs, combining retrieval and In-Context Learning (ICL) techniques to enhance LLM performance (Gao et al., 2023). Classic RAG systems are primarily divided into three modules: the knowledge base module, the retrieval module, and the generation module.

(1) Knowledge Base Module. This module parses, chunks, and vectorizes texts of various formats, storing text chunks and embedding vectors as key-value pairs for rapid and frequent retrieval.

(2) Retrieval Module. This module converts user queries into vectors, matching the top K most similar text chunks from the knowledge base as retrieved contextual information. To improve retrieval accuracy, methods such as Query Rewriting are introduced to reformulate queries for better document relevance (Ma et al., 2023). Sawarkar et al. (2024) introduce both lexical search and semantic search modes during the retrieval process.

(3) Generation Module. This module post-processes the context and ultimately inputs it, along with the question, to the LLMs for response generation. Mao et al. (2024) additionally introduces a document processing module, resulting in a more refined and accurate context. Beyond general domains, RAG is also applied in vertical fields, such as agriculture (Gupta et al., 2024) and finance (Li et al., 2024b).

3 Datasets for Ancient Chinese Poetry

3.1 ACP-Corpus

The proposed ACP-Corpus includes a comprehensive collection of ancient poetry and various types of knowledge texts. We scrape 1,446,096 ancient poems and 46,388 author profiles from the SouYun (Chen, 2009). Each poem is meticulously catalogued in a dictionary format, detailing its unique identifier, title, dynasty, author, genre, and textual content. This corpus also integrates filtered content from other notable sources such as Chinese-poetry (Chinese-Poetry, 2017), Poetry (Werneror, 2017), and Chinese-poetry-and-prose (VMIJUNV, 2022) to create a vast resource library of ancient poetry.

We conduct the following detailed data processing on the ancient poetry resource library. **(1) Handling anomalous characters.** Characters such as "?" and "■" indicate that text has not been displayed correctly. We filter these anomalous characters for manual correction and replacement. Missing Chinese characters in the poetry are uniformly represented by "□". Poems that contain excessive "□" characters will be excluded. **(2) Removing extraneous content.** We utilize regular expressions to match and eliminate unnecessary spaces, consecutive punctuation marks, and irrelevant symbols such as "\r" and "\n". **(3) Information proofreading.** We conduct batch processing and proofreading of titles, dynasties, and author information, including standardizing the expression of titles and dynasties, as well as the consistent representation of authors listed as "unknown". **(4) Deduplication.** We implement a two-stage deduplication process. Stage one involves complete deduplication, where we use the content of the poems as the basis for deduplication, calculating the Hash values of the content and utilizing a Hash set to detect and exclude duplicates. Stage two involves partial deduplication. In the data entries, there may be instances where both a complete ancient poem and a partial excerpt coexist. For these cases, we segment the content using punctuation marks into N text chunks and compare these chunks to filter out partial duplicates. More details of the process can be found in Appendix A.1.3.

Finally, we obtain 1,124,024 ancient poems, contributing from 44,347 authors, spanning 26 dynasties and encapsulating 722 themes. In addition to the ancient poems, we curate another 990,801 entries from the Internet, which include English translations, analyses of ancient poems, vernacular translations, word explanations, idioms, and literary knowledge. As shown in Table 1, the ACP-Corpus demonstrates a significant advantage compared to other corpora.

3.2 ACP-QA

The ACP-QA dataset, derived from the ACP-Corpus and additional web resources, serves as a specialized knowledge base for the RAG system. We focus the instruction dimension on the "knowledge" and "comprehension" levels, disregarding the "generation" level. This is because "generation" type instructions are more suitable for fine-tuning LLMs rather than being part of a retrieval knowledge base. As shown in Table 3, we define 24 different task categories at the "knowledge" and "comprehension" levels and use four different methods to construct the Q&A pairs:

(1) Manual Construction. We directly obtain Q&A pairs through web scraping, manual question creation, and other means. **(2) Template Construction.** Using ERNIE-4.0 (Baidu, 2023) and GPT-4 (OpenAI et al., 2024), we generate question templates for different tasks. After manual

Task	Scale	Proportion
Knowledge		
Content to Title	1,124,024	8.941%
Content to Author	1,124,024	8.941%
Content to Dynasty	1,124,024	8.941%
Content to Three Elements	1,124,024	8.941%
Poem Chain	2,144,524	17.058%
Title to Author	1,124,024	8.941%
Title and Author to Content	1,124,024	8.941%
Poet Introduction	35,523	0.283%
Genre Judgment	780,924	6.212%
Concept Q&A	214	0.002%
Book Introduction	97	0.001%
The Origin of Idiom	5,989	0.048%
Idiom Finding	9,198	0.073%
Poetry Competition	1,644	0.013%
Comprehension		
Vernacular Translation	808,066	6.428%
Poem Appreciation	6,152	0.049%
Word Explanation	83,421	0.664%
English Translation	676	0.005%
Theme Judgment	883,005	7.024%
Imagery Explanation	100,714	0.801%
Appreciation Exam Question	8,850	0.070%
Sentiment Classification	4,000	0.032%
Comprehension Dictation	515	0.004%
Vernacular to Poem	24,498	0.195%
Other		
Other	929,697	7.395%
Overall	12,571,851	100.000%

Table 3: Statistical information of the ACP-QA dataset.

screening and verification, we fill in the blanks in the templates with information to obtain the corresponding Q&A pairs. For example, a template for the task "Title and Author to Content" is: "What is the specific content of [] written by []?". By filling the first [] with the title and the second [] with the author, we obtain a set of questions. To ensure the richness and diversity of the Q&A pairs, we generate a total of 1,121 templates. **(3) LLM Generation.** We use ERNIE-4.0 to segment the knowledge text and guide the LLMs to generate corresponding Q&A pairs based on prompt engineering. **(4) Other Datasets.** We additionally incorporate instructions from other datasets to enrich the knowledge base, including the ACCN-INS and the COIG-CQIA dataset. Ultimately, the scale of ACP-QA reaches 12,571,851 entries. Table 2 demonstrates the superiority of ACP-QA compared to other datasets.

3.3 ACP-Eval

We reconstruct an evaluation dataset called ACP-Eval, based on the tasks in ACP-QA, which contains 7,050 entries. During the reconstruction process, we randomly select relevant knowledge points of ancient poetry and reformulate the questions to minimize overlap with the Q&A pairs in ACP-QA. The primary purpose of ACP-Eval is to evaluate the comprehensiveness of ACP-QA as a RAG knowledge base and to evaluate the retrieval capability of the RAG framework.

4 ACP-RAG

In this section, we present ACP-RAG, a RAG framework specifically designed for the field of ancient Chinese poetry, as illustrated in Figure 1. This framework is designed to be a plug-and-play component that is compatible with exiting high-performing LLMs.

4.1 Retrieval Module

The purpose of the retrieval module is to extract relevant knowledge chunks from the knowledge base based on user queries. To achieve this, we employ a hybrid retrieval approach that integrates both semantic and keyword retrieval pathways to obtain the relevant indices of the knowledge chunks. The semantic retrieval pathway comprises coarse-grained and fine-grained retrieval modules, while the keyword retrieval pathway is facilitated by the keyword retrieval module.

4.1.1 Generate Vector Knowledge Base

Prior to retrieval, the data from the knowledge base should be embedded and stored in the FAISS (Facebook AI Similarity Search) (Johnson et al., 2019) index using a trained embedding model. Each question from the ACP-QA dataset is converted into a 1024-dimensional vector, normalized for consistency, and then batch-stored in the FAISS index, forming a vector knowledge base for retrieval. High-quality ancient Chinese poetry data provides a solid foundation for the framework.

4.1.2 Semantic Coarse-Grained Retrieval Module

This module involves vectorizing the user query into e_{norm} using an embedding model. FAISS then retrieves the top 30 vectors from the offline database by calculating cosine similarity (Equation 1) with e_{norm}, returning the indices of the most similar vectors to achieve semantic coarse-grained retrieval.

$$\text{Cosine Similarity}(e_{\text{norm}}, d_i) = \frac{e_{\text{norm}} \cdot d_i}{\|e_{\text{norm}}\| \|d_i\|} \quad (1)$$

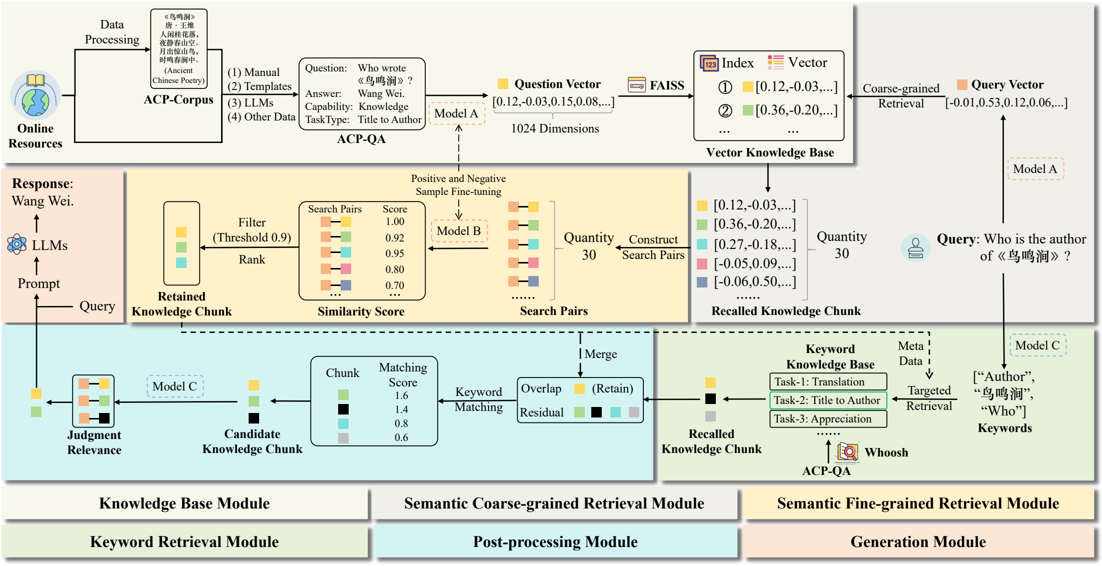

Figure 1: Overview of the ACP-RAG framework. Model A is the embedding model, Model B is the rank model, and Model C is the keyword extraction and context filtering model (Qwen1.5-7B). Zoom in for better view.

4.1.3 Semantic Fine-Grained Retrieval Module

After retrieving the top 30 vectors, the corresponding knowledge chunks are extracted, and user questions are concatenated with each chunk to create query pairs, which are then input into the rank model for similarity scoring. To optimize scoring effectiveness within the 512-token limit for query pairs, we implement the innovative **Truncated Boundary Enhancement (TBE) strategy**. For excessively long texts, TBE retains 100 characters from both the beginning and end to preserve semantic integrity. In contrast, for shorter texts, it repeats content to enrich contextual information. Query pairs that score below a specified threshold are removed, and the remaining pairs are sorted in descending order based on their scores. Consequently, the knowledge chunks that rank higher demonstrate greater relevance to the user questions, effectively minimizing the impact of irrelevant knowledge.

The TBE strategy is particularly suitable for the field of ancient Chinese poetry, primarily because poetry often exhibits strong contextual dependencies and unique expressions. For longer texts, the TBE strategy retains the core elements of the poem's context and imagery, minimizing the loss of important information during truncation. For shorter texts, the repetition of content helps enrich the contextual information, enabling the relevant knowledge chunks to better match the user's query during retrieval.

4.1.4 Keyword Search Module

Unlike traditional keyword search, the keyword search module, building on the previous retrieval steps, consists of three key components: **(1) Knowledge Base Partitioning Strategy.** We utilize Whoosh (Mchaput, 2009) to construct 24 task-specific data indexes tailored to various task types, employing an inverted index mechanism that facilitates rapid keyword retrieval. **(2) Keyword Extraction.** The Qwen 1.5-7B model (Bai et al., 2023) is employed to extract relevant keywords from user queries. **(3) Utilization of Task Information from the Semantic Fine-Grained Retrieval Module.** A distinctive innovation of our approach lies in leveraging metadata associated with each knowledge chunk in ACP-QA, including task type. This allows the module to conduct targeted keyword searches within the appropriate task data indexes based on the task types of the top three scoring knowledge chunks. By narrowing the search scope using this task information, we significantly enhance the relevance and accuracy of the retrieval process. Ultimately, this refined approach yields a set of indexes through efficient keyword searching.

In the above process, semantic retrieval captures the overall meaning of highly condensed imagery, while keyword-based retrieval excels at identifying cultural symbols and fixed expressions. The combination of both complements each other, making it suitable for the field of ancient Chinese poetry.

4.2 Post-Processing Module

The index sequences obtained from the semantic fine-grained retrieval and keyword retrieval modules are merged and processed in the post-processing module to derive the final contextual content. The post-processing module consists of two components: the keyword matching module and the context filtering module.

4.2.1 Keyword Matching Module

After the retrieval process, we introduce an innovative module called the Keyword Matching Module. This module first merges the two sets of index sequences and selects the overlapping indices. The overlapping indices meet both semantic similarity and keyword matching criteria, thus they are retained. The knowledge blocks corresponding to the remaining index are filtered based on keyword matching. For each knowledge chunk *i*, word segmentation is performed using Jieba (Fxsjy, 2012), resulting in the phrase set $N_{(i)}$, and the term frequency matching the query is calculated as $dup_{(i)}$. Additionally, the keyword list N_{query} obtained from the Keyword Retrieval Module is matched with each knowledge chunk *i* to determine the number of matches $match_{(i)}$. Finally, the comprehensive matching score for each knowledge chunk is computed according to Equation 2. If overlapping indices exist, the module retains the top two scoring knowledge chunks; otherwise, it retains the top five. The retained knowledge chunks, along with those corresponding to the overlapping indices, form the candidate knowledge chunks.

$$\text{Score}(i) = \frac{\text{dup}(i)}{N(i)} + \frac{\text{match}(i)}{N_{\text{query}}} \tag{2}$$

Keywords in the field of ancient Chinese poetry are often highly dependent on cultural and contextual factors. This module allows for the further matching and filtering of core imagery and culturally specific terms, thereby enabling more precise retrieval of knowledge chunks that closely align with the query in terms of linguistic characteristics and cultural context.

4.2.2 Context Filtering Module

This module performs the final step of filtering candidate knowledge chunks. Each candidate knowledge chunk is inputted into the Qwen1.5-7B model alongside the question, with the model acting as a "scoring expert". The model scores each chunk based on whether the knowledge chunk aids in answering the question. The scores range from 0, 1, 2, to 3. Knowledge chunks scoring below 2 are discarded. The remaining knowledge chunks are organized into the prescribed format and inputted into the LLMs tasked with answering the question.

4.3 Model Training

4.3.1 Training of the Embedding Model

We use a DSSM (Deep Structured Semantic Model) architecture (Huang et al., 2013) based on BERT (Devlin et al., 2018) to train the embedding model, aiming to map similar ancient poetry texts closer in vector space. As shown in Figure 2, the model structure comprises two BERT models with shared parameters, each generating the vector representation of one sentence in the sample pair. The cosine similarity of the vectors is then calculated and passed through a Sigmoid activation function to produce a similarity score. The similarity score formula is:

$$S = \sigma\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{t\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right) \tag{3}$$

where v_1 and v_2 are the sentence vectors, σ is the Sigmoid function, and *t* is the temperature parameter, which we set to 0.05.

For the base model, we choose BERT-Guwen (Ethan-yt, 2020), which has been pre-trained on a large corpus of classical Chinese texts, as the base model. Regarding data, we employ a template construction method with the aid of ERNIE-4.0 to generate 180,000 positive sample pairs with similar meanings and 500,000 negative sample pairs with different meanings as training data. During training, we fine-tune only the pooler layer of the model, keeping the parameters of other layers frozen to reduce training costs. The loss function used is BCELoss, which is defined as follows:

$$\text{BCELoss}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{4}$$

where y represents the true label set of the batch, and y_i represents the true label of the sample pair (1 for positive, 0 for negative).

4.3.2 Training of the Rank Model

The rank model focuses on scoring and ranking the similarity of retrieved knowledge chunks. We use the same positive and negative sample data for

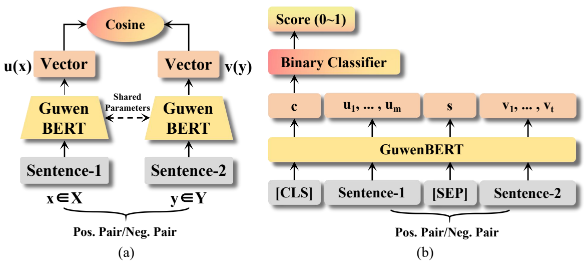

Figure 2: (a) Embedding model training. (b) Rank model training.

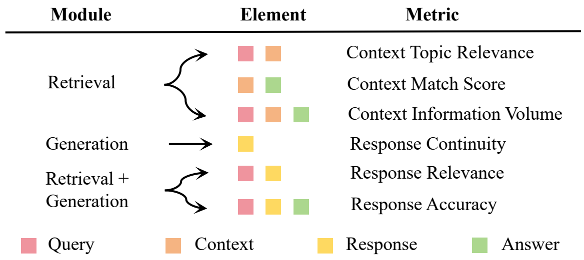

Figure 3: RAG evaluation system.

training. During the training process, we input sentence pairs simultaneously into the BERT-Guwen model for processing. This approach enhances semantic information interaction between sentences, improving scoring accuracy. We fine-tune only the last layer of the BERT model, the pooler layer, and the classification layer, using BCEWithLogitsLoss to optimize the parameters.

5 Experiments

5.1 Baselines

The methods compared in this paper are divided into four categories. **(1) General LLMs**, including GPT-4 (OpenAI et al., 2024), Qwen1.5-7B (Bai et al., 2023), Baichuan2-7B (Yang et al., 2023), and LLaMA3-Chinese-8B (Wang, 2024). **(2) LLMs for classical Chinese**, including Bloom-7B-Chunhua (Wptoux, 2024) and Xunzi1.5 (Shen et al., 2024). We also randomly sample 72,000 Q&A pairs from ACP-QA for fine-tuning Qwen1.5-7B, resulting in Qwen1.5-7B-SFT. **(3) Industrial RAG**, including Kimi (Moonshot, 2023), Perplexity.ai (Perplexity, 2022), and ERNIE-4.0 (Baidu, 2023), all with retrieval functionalities enabled. **(4) Reproducible RAG methods**, including LangChain-ChatChat (Liu et al., 2024a), LLaMAIndex-RAG (Liu, 2022), Self-RAG (Asai et al., 2023), and SAIL (Luo et al., 2023).

5.2 Experimental Settings

(1) Evaluation Dataset. We use ACP-Eval as the evaluation dataset in this paper. **The evaluation of ACP-RAG on other open-source datasets can be found in Appendix C.4.** **(2) Retrieval Sources.** Industry RAG systems retrieve content from the Internet, while the reproducible RAG methods and ACP-RAG retrieve content from the ancient Chinese poetry knowledge base. **(3) RAG Settings.** For ACP-RAG, the coarse-grained retrieval count is set to 30, the fine-grained retrieval threshold is set to 0.9, and the generation module uses the Qwen1.5-7B. Parameters for the reproducible RAG methods are set to the recommended values provided. **(4) Inference Settings.** We use bf16 half-precision inference and greedy decoding strategy, with a maximum generation length set to 1024. The temperature parameter, Top-p, and Top-k sampling are set to 1, 1, and 50, respectively. All experiments are conducted on NVIDIA A6000 GPUs. The results are all averages from the two experiments.

5.3 Experimental Metrics

As shown in Figure 3, based on the relationships among four elements (Query, Context, Response, Answer) and three types of modules, we propose six metrics for evaluating RAG. **(1) Response Accuracy (RA):** the number of correct points in the response. **(2) Response Continuity (RC):** the presence of grammatical errors or content repetition in the response. **(3) Response Relevance (RR):** the relevance of the response to the question. **(4) Context Information Volume (CIV):** the proportion of useful information in the context. **(5) Context Match Score (CMS):** the number of correct points in the answer that match the context. **(6) Context Topic Relevance (CTR):** the relevance of the context to the question. All metrics range from 0% to 100%, with higher values being better.

We specifically fine-tune a scoring model based on Qwen1.5-7B to evaluate the six metrics of different methods using LLM scoring. The scoring results of the model align with human preferences with 92.34% consistency. **For detailed information, please refer to Appendix B.4.**

5.4 Results

The experimental comparison results are presented in Table 4.

Method	RA	RC	RR	CIV	CMS	CTR
General LLMs without Retrieval						
LLaMA3-Chinese-8B	38.6	98.1	77.6	-	-	-
Baichuan2-7B	41.9	98.3	91.6	-	-	-
GPT-4	45.8	99.4	79.0	-	-	-
Qwen1.5-7B	49.2	99.5	84.1	-	-	-
LLMs for Classical Chinese without Retrieval						
Bloom-7B-Chunhua	29.4	95.8	86.2	-	-	-
Xunzi1.5	35.5	98.3	92.8	-	-	-
Qwen1.5-7B-SFT	51.3	98.1	94.8	-	-	-
Industry RAG						
Perplexity.ai	52.2	98.6	94.0	-	-	-
Kimi	75.4	99.5	90.4	-	-	-
ERNIE-4.0	76.8	**99.9**	86.3	-	-	-
Reproducible RAG Methods						
SAIL	36.0	94.9	83.2	21.1	32.5	53.6
LLaMAIndex-RAG	49.8	99.0	89.3	18.0	31.0	51.1
LangChain-ChatChat	56.1	99.1	87.2	40.6	52.6	56.7
Self-RAG	71.5	99.5	92.7	32.0	69.2	82.5
ACP-RAG (Ours)	89.0	99.4	**96.9**	**63.1**	**92.3**	**91.4**
ACP-RAG + SFT (Ours)	**92.4**	98.1	94.4	**63.1**	**92.3**	**91.4**

Table 4: Comparison between ACP-RAG and other methods on ACP-Eval. "RA" indicates Response Accuracy, "RC" indicates Response Continuity, "RR" indicates Response Relevance, "CIV" indicates Context Information Volume, "CMS" indicates Context Match Score, and "CTR" indicates Context Topic Relevance.

5.4.1 Comparison with LLMs without Retrieval

LLMs without retrieval include two types: General LLMs and LLMs for classical Chinese. (1) ACP-RAG significantly outperforms general domain and classical Chinese domain LLMs. The response accuracy of the high-performing models Qwen1.5-7B-SFT and Qwen1.5-7B is only 51.3% and 49.2%, respectively, while ACP-RAG achieves a response accuracy of 89.0%, representing improvements of 37.7% and 39.8%, respectively. The response relevance also increases by 2.1% and 12.8%, respectively. (2) The response continuity of all models is generally above 98%, except for the Bloom-7B-Chunhua model, which still shows a slight deficiency in continuity. (3) When we replace the generation model in ACP-RAG with Qwen1.5-7B-SFT, the response accuracy further improves by 3.4%, indicating that the combination of Supervised Fine-Tuning (SFT) and RAG technologies yields better performance. Additionally, Qwen1.5-7B, after fine-tuning with a small amount of data, shows an improvement of 2.1%, indirectly reflecting the effectiveness of the ACP-QA data.

5.4.2 Comparison with LLMs with Retrieval

LLMs with retrieval include two types: Industrial RAG and Reproducible RAG methods. (1) ACP-RAG shows certain advantages over industry-standard RAG models. In terms of response accuracy, ACP-RAG is 12.2% higher than the best-performing ERNIE-4.0. In terms of response relevance, ACP-RAG is 2.9% higher than the best-performing Perplexity.ai. This is mainly due to ACP-RAG having a very high-quality knowledge base and specialized processing modules. This also reflects that a high-quality knowledge base is one of the key factors for the effectiveness of RAG. (2) Compared to some classic RAG methods, ACP-RAG is more suited for retrieval in the domain of ancient Chinese poetry. ACP-RAG achieves the best results in all metrics except for response continuity. Specifically, ACP-RAG's context information volume is 22.5% higher, context match score is 23.1% higher, and context topic relevance is 8.9% higher. The reason is that, on one hand, knowledge bases constructed in a Q&A format have a higher retrieval recall rate compared to segmenting documents into knowledge chunks. On the other hand, the fine-tuning of embedding and rank models, the use of hybrid retrieval, and the setting of post-processing modules all optimize the retrieval effect.

Module					Metric					
A	B	C	D	E	RA	RC	RR	CIV	CMS	CTR
✔					70.0	99.4	91.8	42.1	69.1	80.0
✔	✔				70.9	99.4	92.0	42.6	69.3	80.3
✔	✔		✔		74.3	99.4	92.9	43.0	72.0	82.5
		✔			62.1	**99.5**	89.8	57.7	34.5	42.9
✔	✔	✔	✔		**89.1**	99.4	**96.9**	54.0	85.4	88.5
✔	✔	✔	✔	✔	89.0	99.4	**96.9**	**63.1**	**92.3**	**91.4**

Table 5: Module effectiveness experiment. A indicates semantic coarse-grained retrieval, B indicates semantic fine-grained retrieval, C indicates keyword retrieval, D indicates keyword matching, and E indicates context filtering.

5.5 Ablation Study

5.5.1 Module Effectiveness

Table 5 presents the results of the ablation study on different modules within ACP-RAG. (1) The introduction of the semantic fine-grained retrieval module brings improvements of 0.9% and 0.2%, respectively. Although the improvements are small, this module preliminarily filters out irrelevant knowledge chunks, reducing interference and the number of operations in subsequent modules, especially when dealing with long contexts. (2) The keyword matching module shows more notable effects, with RA and CMS metrics improving by 3.4% and 2.7%, respectively. (3) While using the keyword retrieval module alone is not effective, significant performance improvements are observed when both semantic and keyword retrieval are used together, followed by post-processing. This is because the tasks

Method	Metric					
	RA	RC	RR	CIV	CMS	CTR
ACP-RAG	**89.1**	99.4	**96.9**	**54.0**	**85.4**	**88.5**
w/o TBE	88.5	99.4	96.8	51.5	83.9	88.1
w/o Prompt Engineering	88.7	**99.5**	96.7	**54.0**	**85.4**	**88.5**

Table 6: Method effectiveness experiment. "w/o" indicates without.

Model	Metric					
	RA	RC	RR	CIV	CMS	CTR
BERT-MLM	69.7	**99.5**	91.2	24.5	43.2	63.3
BERT-Guwen	87.4	99.4	96.8	50.5	81.9	86.9
BERT-Guwen-SFT	**89.1**	99.4	**96.9**	**54.0**	**85.4**	**88.5**

Table 7: Ablation study on the BERT model

in ancient Chinese poetry are diverse; some tasks benefit more from semantic retrieval, while others perform better with keyword retrieval. The combination of both retrieval methods complements each other in task performance. (4) The context filtering module effectively enhances the context metrics, making the final context content more concise and important, with only a 0.1% decrease in response accuracy.

5.5.2 Method Effectiveness

The experimental results in Table 6 show that the TBE strategy in the semantic fine-grained retrieval module improves the retrieval effectiveness for context, demonstrating its effectiveness in ancient Chinese poetry retrieval. Additionally, when the final context and question are input into the generation model, not using a prompt strategy results in a slight decrease in response quality. However, using prompt engineering, where the LLMs act as experts in ancient Chinese poetry and context learning, and provide comprehensive answers based on their knowledge and reference materials, further improves response quality.

5.5.3 BERT Model and Threshold Selection

Table 7 presents the metrics for selecting different embedding models, demonstrating the effectiveness of fine-tuning the model. Table 8 shows the metrics for selecting different fine-grained retrieval thresholds, with 0.9 ultimately chosen as the threshold parameter. Overall, this module is not sensitive to threshold selection, as the metrics show minimal variation and remain relatively stable across different thresholds. This indicates that ACP-RAG maintains stability and robustness, delivering good performance without the need for precise threshold parameter tuning.

Threshold	Metric					
	RA	RC	RR	CIV	CMS	CTR
0.8	88.6	**99.4**	**97.3**	51.0	83.4	87.7
0.9	**89.1**	**99.4**	96.9	**54.0**	**85.4**	**88.5**
0.95	88.7	**99.4**	97.0	51.0	83.4	87.7
0.99	88.6	**99.4**	97.0	51.1	83.4	87.6

Table 8: Ablation study on threshold selection in the semantic fine-grained retrieval module.

6 Conclusion

In this paper, we presents a comprehensive framework, ACP-RAG, for enhancing the performance of Large Language Models (LLMs) in the domain of ancient Chinese poetry. The ACP-RAG framework incorporates multiple modules, including semantic retrieval, keyword retrieval, and context filtering, to improve the accuracy and relevance of LLM responses. Moreover, through the construction of three new datasets, namely the ACP-Corpus, ACP-QA, and ACP-Eval, we address the challenges of data scarcity and inadequate cultural context in existing LLMs. Experimental results demonstrate significant improvements of our method compared to existing LLMs. It is our hope that this work will contribute to the inheritance and innovative development of ancient poetry culture, paving the way for future research in preserving and advancing cultural heritage within the context of LLMs.

7 Limitations

This study has made some progress in enhancing the ability of LLMs to process ancient Chinese poetry, but there are still some limitations. Firstly, the datasets may contain historical background biases, and the model may unintentionally reinforce these biases. Secondly, although the system performs well in the Chinese context, its cross-cultural or cross-lingual applicability has not been fully verified.

Acknowledgements

This research is supported in part by National Natural Science Foundation of China (Grant No.: 62441604, 62476093).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Self-reflective retrieval augmented generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. 2024. COIG-CQIA: Quality is all you need for Chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*.

Baidu. 2023. ERNIE. https://yiyan.baidu.com/.

Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024a. WenMind: A comprehensive benchmark for evaluating large language models in Chinese classical literature and language arts. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024b. TongGu: Mastering classical Chinese understanding with knowledge-grounded large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4196–4210, Miami, Florida, USA. Association for Computational Linguistics.

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

Yiyun Chen. 2009. SouYun. https://www.sou-yun.cn/.

Chinese-Poetry. 2017. Chinese-poetry. https://github.-com/chinese-poetry/chinese-poetry. GitHub.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Ethan-yt. 2020. GuwenBERT. https://github.com/Ethan-yt/guwenbert. GitHub.

FlagAlpha. 2024. LLaMA2-Chinese-7B-Chat. https://-huggingface.co/FlagAlpha/Llama2-Chinese-7b-Chat. Hugging Face.

Fxsjy. 2012. Jieba. https://github.com/fxsjy/jieba. GitHub.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, et al. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18099–18107.

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

Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li, Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, Morris Sharp, et al. 2024. RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*.

Jing He, M. Zhou, and Long Jiang. 2012. Generating Chinese classical poems with statistical machine translation models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2333–2338, New York, NY, USA. Association for Computing Machinery.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Iflytek. 2023. Spark-v3.5. https://xinghuo.xfyun.cn/.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. CCPM: A Chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, Jun Huang, and Wei Lin. 2024b. AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. *arXiv preprint arXiv:2403.12582*.

Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Noah Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, Wangchunshu Zhou, Yiming Liang, Lei Zhang, Lei Ma, Jiajun Zhang, Zuowen Li, Wenhao Huang, Chenghua Lin, and Jie Fu. 2024c. CIF-Bench: A Chinese instruction-following benchmark for evaluating the generalizability of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12431–12446, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Liu. 2022. LLaMAIndex. https://github.com/-jerryjliu/llama_index. Github.

Qian Liu, Jinke Song, Zhiguo Huang, and Yuxuan Zhang. 2024a. LangChain-ChatChat. https://github.com/chatchat-space/Langchain-Chatchat. Github.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024b. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*.

Hongyin Luo, Tianhua Zhang, Yung-Sung Chuang, Yuan Gong, Yoon Kim, Xixin Wu, Helen Meng, and James Glass. 2023. Search augmented instruction learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3717–3729.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Yuren Mao, Xuemei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. 2024. FIT-RAG: Black-box RAG with factual information and token reduction. *arXiv preprint arXiv:2403.14374*.

Mchaput. 2009. Whoosh. https://github.com/mchaput/whoosh. GitHub.

Moonshot. 2023. Kimi.AI. https://kimi.moonshot.cn/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. GPT-4 technical report.

Perplexity. 2022. Perplexity.AI. https://www.perplexity.ai/.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended RAG: Improving RAG (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. *arXiv preprint arXiv:2404.07220*.

Si Shen, Bing Li, and Xueliang Ma. 2024. XunziALLM. https://github.com/Xunzi-LLM-of-Chinese-classics/Xunzi-ALLM. GitHub.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. In *In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106, Barcelona, Spain (Online). Association for Computational Linguistics.

Truera. 2023. Trulens. https://github.com/truera/trulens. GitHub.

VMIJUNV. 2022. Chinese-poetry-and-prose. https://github.com/VMIJUNV/Chinese-poetry-and-prose. GitHub.

Shenzhi Wang. 2024. LLaMA3-8B-Chinese-Chat. https://huggingface.co/shenzhiwang/Llama3-8B-Chinese-Chat. Hugging Face.

Werneror. 2017. Poetry. https://github.com/Werneror/Poetry. GitHub.

Wptoux. 2024. Bloom-7B-Chunhua. https://huggingface.co/wptoux/bloom-7b-chunhua. Hugging Face.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jianxin Yang. 2023. Firefly. https://github.com/yangjianxin1/Firefly. GitHub.

Xiaoyuan Yi, Ruoyu Li, Cheng Yang, Wenhao Li, and Maosong Sun. 2020. MixPoet: Diverse poetry generation via learning controllable mixed latent space. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence*, New York, USA.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01.AI.

Ming Zhang, Yue Zhang, Shichun Liu, Haipeng Yuan, Junzhe Wang, Yurui Dong, Jingyi Deng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023a. LLMEval-2. https://github.com/llmeval/llmeval-2.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. Evaluating the performance of large language models on GAOKAO benchmark. *arXiv preprint arXiv:2305.12474*.

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

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

A Data

A.1 ACP-Corpus

A.1.1 Distribution of Dynasties in Ancient Chinese Poetry Corpora

Table 9 presents the distribution of ancient Chinese poems and authors across different dynasties in the THU-CCPC (Guo et al., 2019), Chinese-poetry (Chinese-Poetry, 2017), Poetry (Werneror, 2017), Chinese-poetry-and-prose (VMIJUNV, 2022), and ACP-Corpus (Ours) corpora. It is evident that the ACP-Corpus covers 26 different dynasties, making it the most extensive among all the corpora. Additionally, the ACP-Corpus contains the highest number of ancient poems and authors.

A.1.2 Examples of Ancient Chinese Poetry

Figure 4 illustrates the ancient Chinese poetry data in the ACP-Corpus. Each piece of poetry is stored in the form of a dictionary, containing 7 key-value pairs: ID, Title, Dynasty, Author, Kind, Content, and Content_split. Here, "Kind" refers to the genre of the poetry, and "Content_split" stores each sentence of the poetry in a list format after splitting the content based on punctuation marks.


```
{
  "Id": 74590,
  "Title": "题刘偃庄",
  "Dynasty": "唐",
  "Author": "刘商",
  "Kind": "七言绝句",
  "Content": "何事退耕沧海畔，闲看富贵白云飞。门前种稻三回熟，县里官人四考归。",
  "Content_split": [
    "何事退耕沧海畔，",
    "闲看富贵白云飞。",
    "门前种稻三回熟，",
    "县里官人四考归。"
  ]
}
```

```
{
  "Id": 1066079,
  "Title": "偷声木兰花·春分遇雨",
  "Dynasty": "宋",
  "Author": "徐铉",
  "Kind": "词",
  "Content": "天将小雨交春半，谁见枝头花历乱。纵目天涯，浅黛春山处处纱。焦人不过轻寒恼，问卜怕听情未了。许是今生，误把前生草踏青。",
  "Content_split": [
    "天将小雨交春半，",
    "谁见枝头花历乱。",
    "纵目天涯，",
    "浅黛春山处处纱。",
    "焦人不过轻寒恼，",
    "问卜怕听情未了。",
    "许是今生，",
    "误把前生草踏青。"
  ]
}
```


Figure 4: Original examples of ancient Chinese poetry in the ACP-Corpus.

A.1.3 Deduplication of Ancient Chinese Poetry

For ancient Chinese poetry data, we implement a two-stage deduplication process. Stage one involves exact deduplication, while stage two addresses partial deduplication.

Stage One: Exact deduplication is performed using a Hash algorithm. The Hash algorithm converts data of arbitrary length into a fixed-length Hash value through a specific computation method, with the resulting Hash values generally being unique.

Algorithm 1 Duplicate Poem Detection

Require: Poem P_s (shorter poem)
Require: Poem P_l (longer poem)
Ensure: Keep the longer poem if the poems are duplicates

$N_s \leftarrow$ number of sentence blocks in P_s
$N_l \leftarrow$ number of sentence blocks in P_l
$C \leftarrow$ number of common sentence blocks between P_s and P_l
if $C > \frac{N_s}{2}$ **then**
    Keep P_l
end if

By calculating the Hash values of the poetry content items, data items with identical Hash values can be removed, efficiently eliminating duplicates from the dataset.

Stage Two: In the collected ancient poetry data, partial duplicates may occur. For example, some data may split the same poem into multiple segments, treating each segment as an independent entity, which results in partial duplicates of the whole poem and its segmented parts. The Hash algorithm alone cannot accurately identify and remove these partial duplicates. To address this, we compare the original texts of the poems to filter out partial duplicates. Specifically, if more than half of the content of a shorter poem matches the content of a longer poem, the two poems are considered duplicates, with the longer poem retained. This process follows the procedure outlined in Algorithm 1.

A.2 ACP-QA & ACP-Eval

A.2.1 Task Information

Table 10 and 11 provides a detailed description of the 24 types of ancient poetry-related tasks in ACP-QA and ACP-Eval. The table includes the following information: task number, task names in English, task description, task dimension, and the average token length of Q&A pairs for each task in ACP-QA and ACP-Eval.

A.2.2 Task Source

Table 12 presents the data sources corresponding to the 24 types of ancient Chinese poetry tasks, including specific data sources, relevant links, and the licenses adhered to by the source data. It is important to emphasize that both the ACP-QA and ACP-Eval datasets also follow the licenses of the original data.

Dynasties	THU-CCPC		Chinese-Poetry		Poetry		Chinese-poetry-and-prose		ACP-Corpus	
	# Poems	**# Authors**	**# Poems**	**# Authors**	**# Poems**	**# Authors**	**# Poems**	**# Authors**	**# Poems**	**# Authors**
Ancient Times	0	0	0	0	0	0	0	0	2	2
Xia	0	0	0	0	0	0	0	0	1	1
Shang	0	0	0	0	0	0	0	0	3	3
Zhou	0	0	0	0	0	0	553	15	9	5
Spring and Autumn	0	0	305	1	0	0	0	0	309	5
Warring States	0	0	0	0	0	0	0	0	55	6
Pre-Qin	0	0	0	0	570	8	0	0	240	6
Qin	0	0	0	0	2	2	9	5	15	6
Han	0	0	65	10	363	83	551	101	1,244	137
Three Kingdoms	0	0	26	1	0	0	419	32	688	63
Jin (266-420 AD)	0	0	0	0	3,020	251	1,811	214	3,065	294
Sixteen Kingdoms	0	0	0	0	0	0	0	0	22	8
Southern and Northern	3	3	0	0	4,587	435	4,250	437	4,363	503
Sui	79	19	0	0	1,170	84	1,353	122	907	88
Tang	9,361	1,231	107,891	7,983	49,667	2,776	49,149	3,125	50,690	3,269
Wu Zhou	0	0	0	0	0	0	0	0	190	35
Five Dynasties and Ten Kingdoms	0	0	542	20	0	0	2,874	207	5,445	434
Song	58,964	3,368	275,807	10,539	288,232	9,490	257,418	9,472	253,275	9,712
Liao	4	2	0	0	22	7	24	10	24	10
Jin (1115-1234 AD)	1,513	194	0	0	2,975	262	4,180	309	8,186	371
Yuan	9,214	639	11,057	233	52,452	1,267	61,013	1,964	79,404	2,512
Ming	48,539	2,651	0	0	252,693	4,518	257,865	8,537	309,102	9,210
Qing	5	4	477	2	107,789	9,048	193,586	13,832	283,759	14,243
Republic of China	0	0	0	0	15,367	99	0	0	45,546	432
Modern	0	0	0	0	12,464	48	0	0	10,922	2,333
Contemporary	0	0	0	0	1,948	9	509	36	66,558	749
Overall	127,682	8,111	396,170	18,789	793,321	28,387	835,564	38,418	**1,124,024**	**44,437**

Table 9: Comparison of the distribution of dynasties in the ancient Chinese poetry corpora.

ID	Task Name	Task Description
T1	Content to Title	Answer the title of the ancient poem based on the content
T2	Content to Author	Answer the author of the ancient poem based on the content
T3	Content to Dynasty	Answer the dynasty of the ancient poem based on the content
T4	Content to Three Elements	Answer the title, author, and dynasty of the ancient poem based on the content (three elements)
T5	Vernacular Translation	Translate the ancient poem into modern vernacular Chinese
T6	Poem Appreciation	Conduct a free appreciation of the ancient poem
T7	Word Explanation	Explain the meanings of words in the ancient poem
T8	Poem Chain	Answer the second half (first half) of the ancient poem based on the first half (second half)
T9	English Translation	Translate the ancient poem into English
T10	Title to Author	Answer the author of the ancient poem based on the title
T11	Title and Author to Content	Answer the content of the ancient poem based on the title and author
T12	Poet Introduction	Provide a brief introduction to the poet
T13	Genre Judgment	Judge the genre of the ancient poem
T14	Theme Judgment	Judge the theme of the ancient poem
T15	Imagery Explanation	Explain the meanings of the imagery present in the ancient poem
T16	Concept Q&A	Provide the meanings of concepts related to ancient poetry
T17	Book Introduction	Provide a brief introduction to books related to ancient poetry
T18	The Origin of Idiom	Provide the origin and meaning of the idiom
T19	Idiom Finding	Identify the idiom contained in the ancient poem and provide its meaning
T20	Appreciation Exam Question	Real exam questions on the appreciation of the ancient poem
T21	Comprehension Dictation	Provide the corresponding lines of the ancient poem based on the relevant prompts
T22	Poetry Competition	Real questions from competitions and programs related to ancient poetry
T23	Sentiment Classification	Classify the sentiment expressed in the ancient poem
T24	Vernacular to Poem	Retrieve the corresponding ancient poem based on the modern vernacular Chinese

Table 10: Description of 24 ancient Chinese poetry tasks.

A.2.3 Task Examples

Figure 5, 6, 7, 8, 9, and 10 display examples of Q&A pairs for the 24 types of tasks.

A.2.4 Examples of Templates for Constructing Q&A Pairs

We use a template-based method to construct Q&A pairs. Figure 11 presents examples of the templates used to construct Q&A pairs for the Content to Title, Vernacular Translation, and Poet Introduction tasks. In summary, based on different question formats for each task, a set of task templates is generated using LLMs and manual screening. By inserting the metadata of the questions into the corresponding templates, various question and answer formats are produced.

A.2.5 Comparison of ACP-Eval with Other Evaluation Datasets

As shown in Table 13, we compare the ACP-Eval evaluation dataset used in this paper with other open-source evaluation datasets. Among evaluation datasets for the field of ancient Chinese poetry, ACP-Eval is currently the largest in scale and encompasses the most tasks. It comprehensively examines different dimensions of knowledge and abil-

任务1：内容->题目
Task1： Content to Title

问题：请问下面出自哪首古诗词？两脚著地此何关，白云与尔同去还。正当海阔天高处，不离区区跬步间。
Question: Could you please tell me which ancient poem the following lines are from? 两脚著地此何关，白云与尔同去还。正当海阔天高处，不离区区跬步间。

答案：《次韵张廷实见寄 其一》
Answer: 《次韵张廷实见寄 其一》

任务2：内容->作者
Task2： Content to Author

问题："恍惚之中寻有象，杳冥之内觅真精。有无从此自相入，未见如何想得成。"你能告诉我这首诗的作者吗？
Question: "恍惚之中寻有象，杳冥之内觅真精。有无从此自相入，未见如何想得成。" Can you tell me the author of this poem?

答案：张伯端
Answer: Zhang Boduan.

任务3：内容->朝代
Task3： Content to Dynasty

问题：你了解"一庭芳草堕残红，消尽春光满眼中。大半东皇妒艳色，莫教零落怨东风。"这首诗是哪个朝代的吗？
Question: Do you know which dynasty this poem "一庭芳草堕残红，消尽春光满眼中。大半东皇妒艳色，莫教零落怨东风。" is from?

答案：这首诗是"清"时期的。
Answer: This poem is from the Qing dynasty.

任务4：内容->三要素
Task4： Content to Three Elements

问题：请说出"月浸千林万壑秋，听琴惊起鹤回头。孤云带湿藏空谷，不逐人间富贵浮。"的题目、作者以及朝代。
Question: Please provide the title, author, and dynasty of the poem "月浸千林万壑秋，听琴惊起鹤回头。孤云带湿藏空谷，不逐人间富贵浮。"

答案：题目：《云谷》；作者：徐集孙；朝代/时期：南宋。
Answer: Title: 《云谷》; Author: Xu Jisun; Dynasty/Period: Southern Song.

Figure 5: Examples of tasks 1 through 4.

任务5：古文->白话文
Task5： Vernacular Translation

问题：请帮我翻译诗句“腊底春前探物华，一枝初发竞相誇。”
Question: Please help me translate the poem into modern Chinese: “腊底春前探物华，一枝初发竞相誇。”

答案：腊月初春前，探究了万物的精华，一枝初开的梅花争相夸耀。
Answer: 腊月初春前，探究了万物的精华，一枝初开的梅花争相夸耀。

任务6：古诗词赏析
Task6： Poem Appreciation

问题：题目：登楼 作者：杜甫 朝代：唐 内容：花近高楼伤客心，万方多难此登临。锦江春色来天地，玉垒浮云变古今。北极朝廷终不改，西山寇盗莫相侵。可怜后主还祠庙，日暮聊为梁甫吟。请对这首古诗词进行赏析。
Question: Title: Climbing the Tower Author: Du Fu Dynasty: Tang Content: 花近高楼伤客心，万方多难此登临。锦江春色来天地，玉垒浮云变古今。北极朝廷终不改，西山寇盗莫相侵。可怜后主还祠庙，日暮聊为梁甫吟。Please analyze this ancient poem.

答案：此诗是一首感时抚事的诗。作者写登楼望见无边春色，想到万方多难，浮云变幻，不免伤心感喟......
Answer: This poem is one that reflects on the times and events. The author writes about climbing a tower and seeing the boundless spring scenery, which makes him think of the many troubles across the land and the ever-changing clouds, inevitably feeling saddened and lamenting......

任务7：词语解释
Task7： Word Explanation

问题：内容：欲减罗衣寒未去。不卷珠帘，人在深深处。红杏枝头花几许？啼痕止恨清明雨。尽日沉烟香一缕。宿酒醒迟，恼破春情绪。飞燕又将归信误，小屏风上西江路。“沉烟”具体是什么意思?
Question: Content: 欲减罗衣寒未去。不卷珠帘，人在深深处。红杏枝头花几许？啼痕止恨清明雨。尽日沉烟香一缕。宿酒醒迟，恼破春情绪。飞燕又将归信误，小屏风上西江路。What does “沉烟” specifically mean?

答案：点燃的沉香，常用以营造香气或烘托气氛。
Answer: The ignited agarwood is often used to create fragrance or to enhance the atmosphere.

任务8：古诗词接龙
Task8： Poem Chain

问题：郭印写的《三伏不热》中，“连旬雨复阴”的前一句是哪句?
Question: In Guo Yin’s “Three Hottest Days Are Not Hot” what is the line before “连旬雨复阴”?

答案：三伏元无暑。
Answer: 三伏元无暑。

Figure 6: Examples of tasks 5 through 8.

任务9：古文->英文
Task9： English Translation

问题：请把古诗词“古庙依青嶂，行宫枕碧流。水声山色锁妆楼。”翻译成英文。
Question: Please translate the ancient poem “古庙依青嶂，行宫枕碧流。水声山色锁妆楼。” into English.

答案：Olden temple sits beside jade mountain. Makeshift palace overlooks green river. Water and hill hug concubine residence.
Answer: Olden temple sits beside jade mountain. Makeshift palace overlooks green river. Water and hill hug concubine residence.

任务10：题目->作者
Task10： Title to Author

问题：《和伯阳池上梅花》的作者是谁?
Question: Who is the author of “Responding to Bo Yang: Plum Blossoms on the Pond”?

答案：诗人杨士奇。
Answer: The poet Yang Shiqi.

任务11：题目+作者->内容
Task11：Title and Author to Content

问题：我想了解一下释妙声写的《宽上人云泉》的内容。
Question: I would like to know about the content of “Kuan Shangren Yunquan” written by Shi Miaosheng.

答案：内容：白云在天白于鹤，下映流泉相喷薄......
Answer: Content: 白云在天白于鹤，下映流泉相喷薄......

任务12：介绍人物
Task12： Poet Introduction

问题：介绍一下诗人赵彦肃，人物所属时期为：宋。
Question: Introduce the poet Zhao Yansu, who belonged to the Song dynasty.

答案：宋严州建德人，字子钦，号复斋。私淑陆九渊。孝宗乾道二年进士......
Answer: A native of Jiande in Yan Prefecture during the Song dynasty, with the courtesy name Ziqin and the pseudonym Fuzhai. He privately studied under Lu Jiuyuan. He passed the imperial examination in the second year of Emperor Xiaozong's Qiandao reign......

Figure 7: Examples of tasks 9 through 12.

任务13：判断体裁
Task13： Genre Judgment

问题：绿云连地软，红雨接天飞。一尊谁共赏，满目尽吾诗。请问这首古诗词在体裁上属于哪一种?
Question: 绿云连地软，红雨接天飞。一尊谁共赏，满目尽吾诗。May I ask what genre this ancient poem belongs to?

答案：五言绝句。
Answer: Five-character quatrain.

任务14：判断题材
Task14： Theme Judgment

问题：内容：好是清凉地，都无系绊身。晚晴宜野寺，秋景属闲人...... 这首古诗词主要表达哪种题材?
Question: Content: 好是清凉地，都无系绊身。晚晴宜野寺，秋景属闲人...... What is the main theme expressed in this ancient poem?

答案：水天一色。
Answer: Water and sky of one color.

任务15：意象解释
Task15： Imagery Explanation

问题：请解释“博士”这一意象在古诗词中的象征意义。
Question: Please explain the symbolic meaning of the image of “博士” in ancient poetry.

答案：（1）博通古今的人。（2）古代学官名。（3）对具有某种技艺或专门从事某种职业的人的尊称。（4）学位名，学位的最高一级。
Answer: (1) A person well-versed in ancient and modern knowledge. (2) An ancient academic official title. (3) A respectful title for those with certain skills or who engage in specific professions. (4) An academic degree, the highest level of academic qualification.

任务16：古诗词概念问答
Task16： Concept Q&A

问题：什么是“失韵”？
Question: What is “失韵” ?

答案：即出韵。指作诗词押韵违反格律，使用非同韵部的字。多就近体诗而言。唐宋以来，律赋、骈文、诏令、表启之类，上下句平仄不调，也叫失韵。
Answer: “失韵” refers to breaking rhyme. It indicates that in composing poetry or lyrics, the rhyming violates the metrical rules by using characters that do not belong to the same rhyme category. This term is mostly used in reference to near-form poems. Since the Tang and Song dynasties, it has also referred to mismatches in tone patterns between successive lines in regulated verse, parallel prose, official decrees, memorials, and the like.

Figure 8: Examples of tasks 13 through 16.

任务17：书籍介绍
Task17： Book Introduction

问题：能否用几句话来介绍《花庵词选》这本书?
Question: Could you introduce the book "Selection of Hua'an Lyrics" in a few sentences?

答案：《花庵词选》，南宋黄升编。全书二十卷。《花庵词选》书成于淳祐己酉，前十卷选唐宋诸贤之词，始于李白，终于北宋王昴......
Answer: "Selection of Hua'an Lyrics" compiled by Huang Sheng of the Southern Song dynasty. The entire book consists of twenty volumes. It was completed in the Jiyou year of the Chunyou era. The first ten volumes feature lyrics by various esteemed poets of the Tang and Song dynasties, starting with Li Bai and ending with Wang Ao of the Northern Song dynasty...

任务18：成语问答A-来源
Task18： The Origin of Idiom

问题：给出"本来面目"这个成语的出处和它代表的意义。
Question: Provide the origin and the meaning of the idiom "本来面目".

答案：意思是：原是佛教用语，指人的本性；后比喻人或事物原来的样子。成语出处：宋·苏轼《老人行》："一任秋霜换鬓毛，本来面目常如故。"
Answer: The meaning is: originally a Buddhist term referring to a person's true nature; later used as a metaphor for the original state of a person or thing. Origin of the idiom: Song dynasty, Su Shi, "Old Man's Walk" : "一任秋霜换鬓毛，本来面目常如故。"

任务19：成语问答B-蕴含
Task19： Idiom Finding

问题：不争你握雨携云，常使我提心在口。能否从上文中找出隐含的成语，并解释其含义?
Question: 不争你握雨携云，常使我提心在口。Can you identify the implied idiom from the above text and explain its meaning?

答案：成语：提心在口 释义：心在口边，几乎要跳出来。形容担心、恐惧。
Answer: Idiom: "提心在口" Meaning: The heart is at the edge of the mouth, almost jumping out. It describes a feeling of worry or fear.

任务20：古诗词赏析真题
Task20： Appreciation Exam Question

问题：王安石《壬辰寒食》请简要分析诗的前两联如何表现作者的思乡之情和哀伤情绪。
Question: Please briefly analyze how the first two couplets of Wang Anshi's "Cold Food Festival of the Renchen Year" express the author's homesickness and sorrowful emotions.

答案：通过比喻写客思之愁，用春风吹拂的千万条柳枝来写愁绪之多；通过夸张写寒食之哀，用倾泪可涨冶城之潮写出悲情之重。
Answer: Through metaphor, the poem expresses the sorrow of being a traveler by comparing the countless willow branches swayed by the spring breeze to the abundance of melancholy feelings. Through exaggeration, it conveys the sadness of the Cold Food Festival by describing tears flowing enough to raise the tide of Yecheng, emphasizing the depth of sorrow.

Figure 9: Examples of tasks 17 through 20.

任务21：古诗词理解性默写
Task21： Comprehension Dictation

问题：《离骚》中写自己虽崇尚美德约束自己，多少年仍然遭到贬黜的两句是?
Question: Which two lines in “Li Sao” describe the author’s enduring commitment to virtue and self-restraint despite being dismissed for many years?

答案：余虽好修姱以鞿羁兮，謇朝谇而夕替。
Answer: 余虽好修姱以鞿羁兮，謇朝谇而夕替。

任务22：诗词竞赛真题
Task22： Poetry Competition

问题：中国的文学四大体裁分别是什么?
Question: "What are the four major genres of Chinese literature?

答案：诗歌、小说、散文、戏剧。
Answer: Poetry, novels, prose, drama

任务23：古诗词情感分类
Task23： Sentiment Classification

问题：“生别犹疑不再逢，楚天云树隔重重。愁来读尽荆南稿，风雨空斋掩暮钟。”上述古诗词的情感是“正面”、“负面”还是“中性”的?
Question: “生别犹疑不再逢，楚天云树隔重重。愁来读尽荆南稿，风雨空斋掩暮钟。”
Are the emotions expressed in the above ancient poems “positive” “negative” or “neutral”?

答案：负面
Answer: Negative.

任务24：白话文->古诗词
Task24： Vernacular to Poem

问题：找出下面白话文对应的古诗文。白话文：许多禽鸟大声喧嚷它却独自凝眸沉寂。
Question: Find the ancient Chinese poem corresponding to the following vernacular text: 许多禽鸟大声喧嚷它却独自凝眸沉寂。

答案：众禽喧呼独凝寂。
Answer: 众禽喧呼独凝寂。

Figure 10: Examples of tasks 21 through 24.

ID	Task Name	Capability	ACP-QA		ACP-Eval	
			Avg.Q Tokens	Avg.A Tokens	Avg.Q Tokens	Avg.A Tokens
T1	Content to Title	Knowledge	52.64	13.33	47.12	12.80
T2	Content to Author	Knowledge	53.00	13.55	46.19	13.45
T3	Content to Dynasty	Knowledge	55.61	8.17	46.11	8.20
T4	Content to Three Elements	Knowledge	56.48	20.77	48.04	20.36
T5	Vernacular Translation	Comprehension	47.27	67.99	45.78	65.03
T6	Poem Appreciation	Comprehension	100.78	231.49	83.27	104.60
T7	Word Explanation	Comprehension	173.63	6.68	106.46	6.01
T8	Poem Chain	Knowledge	26.64	3.62	26.01	3.63
T9	English Translation	Comprehension	56.86	192.64	42.37	134.09
T10	Title to Author	Knowledge	15.25	4.25	13.95	4.21
T11	Title and Author to Content	Knowledge	20.76	57.86	19.74	47.92
T12	Poet Introduction	Knowledge	16.72	133.28	17.38	158.08
T13	Genre Judgment	Knowledge	44.65	1.02	51.22	1.14
T14	Theme Judgment	Comprehension	54.19	4.64	46.48	8.39
T15	Imagery Explanation	Comprehension	12.78	135.32	12.46	140.75
T16	Concept Q&A	Knowledge	10.07	105.70	12.79	93.34
T17	Book Introduction	Knowledge	12.94	319.42	13.14	329.76
T18	The Origin of Idiom	Knowledge	15.20	108.12	15.34	104.41
T19	Idiom Finding	Knowledge	35.73	53.58	29.23	35.66
T20	Appreciation Exam Question	Comprehension	102.07	54.32	112.98	54.25
T21	Comprehension Dictation	Comprehension	30.31	9.30	30.71	9.00
T22	Poetry Competition	Knowledge	56.70	1.93	62.54	2.58
T23	Sentiment Classification	Comprehension	82.80	1.61	36.43	1.00
T24	Vernacular to Poem	Comprehension	49.71	1.00	17.58	3.61

Table 11: Relevant information on 24 ancient Chinese poetry tasks. "Avg.Q Tokens" indicates the average token length of the questions, "Avg.A Tokens" indicates the average token length of the answers.

ID	Task Name	Data Source	Related Link	License
T1	Content to Title	Internet	https://www.sou-yun.cn/	CC0 1.0
T2	Content to Author	Internet	https://www.sou-yun.cn/	CC0 1.0
T3	Content to Dynasty	Internet	https://www.sou-yun.cn/	CC0 1.0
T4	Content to Three Elements	Internet	https://www.sou-yun.cn/	CC0 1.0
T5	Vernacular Translation	Internet Chinese-poetry-and-prose	https://www.sou-yun.cn/ https://github.com/VMIJUNV/chinese-poetry-and-prose	CC0 1.0 Open Source
T6	Poem Appreciation	Internet Chinese-poetry-and-prose	https://www.gushixuexi.com/ https://github.com/VMIJUNV/chinese-poetry-and-prose	Open Source Open Source
T7	Word Explanation	Internet Chinese-poetry-and-prose LLMs	https://www.gushixuexi.com/ https://github.com/VMIJUNV/chinese-poetry-and-prose -	Open Source Open Source -
T8	Poem Chain	Internet	https://www.sou-yun.cn/	CC0 1.0
T9	English Translation	Internet	https://www.zhihu.com/	Zhihu User Agreement
T10	Title to Author	Internet	https://www.sou-yun.cn/	CC0 1.0
T11	Title and Author to Content	Internet	https://www.sou-yun.cn/	CC0 1.0
T12	Poet Introduction	Internet	https://www.sou-yun.cn/	CC0 1.0
T13	Genre Judgment	Internet	https://www.sou-yun.cn/	CC0 1.0
T14	Theme Judgment	Internet	https://www.sou-yun.cn/	CC0 1.0
T15	Imagery Explanation	Internet	https://www.sou-yun.cn/	CC0 1.0
T16	Concept Q&A	Internet	-	-
T17	Book Introduction	Internet	Baidu baike	Baidu baike User Agreement
T18	The Origin of Idiom	ACCN-INS Internet	- https://www.hanyuguoxue.com/chengyu/	- Open Source
T19	Idiom Finding	ACCN-INS Internet	- https://www.hanyuguoxue.com/chengyu/	- Open Source
T20	Appreciation Exam Question	Internet Internet	http://ts300.5156edu.com/ http://www.exam58.com/	Open Source Open Source
T21	Comprehension Dictation	Internet	-	-
T22	Poetry Competition	Internet COIG-CQIA	- https://huggingface.co/datasets/m-a-p/COIG-CQIA	- Open Source
T23	Sentiment Classification	Internet LLMs	- -	- -
T24	Vernacular to Poem	Internet LLMs	https://www.sou-yun.cn/ -	CC0 1.0 -

Table 12: Data sources corresponding to the 24 types of tasks

ities related to ancient Chinese poetry, including transcription, translation, appreciation, and general knowledge. The evaluation method for ACP-Eval involves LLMs simulating teachers and scoring based on scoring points to obtain relevant metrics. Therefore, we fine-tune a scoring model specifically for the evaluation of this dataset, as detailed in Appendix B.

任务1: 内容->题目　Task1: Content to Title

"[ ]" 这首诗的题目是什么?
"[ ]" What is the title of this poem ?

"[ ]" 出自哪首古诗词?
Which poem is "[ ]" from ?

告诉我"[ ]"来自哪首诗。
Tell me which poem "[ ]" is from.

任务5: 古文->白话文　Task5: Vernacular Translation

请将诗句"[ ]"翻译成白话文。
Please translate the poem "[ ]" into vernacular Chinese.

诗句"[ ]"的白话文表述是怎样的?
What is the vernacular Chinese expression of the poem "[ ]" ?

将"[ ]"转换成白话文。
Convert "[ ]" into vernacular Chinese.

任务12: 介绍人物　Task12: Poet Introduction

简单介绍一下[ ]。
Briefly introduce [ ].

能否告诉我关于[ ]的基本信息?
Could you tell me the basic information about [ ] ?

请对[ ]进行一番介绍。
Please provide an introduction to [ ].

Figure 11: Examples of task templates used for constructing Q&A pairs

Dataset	Domain	License	Scale	# Tasks	Method		
					HG	CI	MC
C-Eval (Huang et al., 2024)	General	CC BY-NC-SA-4.0	6	1	✔	✘	✔
CIF-Bench (Li et al., 2024c)	General	-	85	2	✔	✔	✘
CMMLU (Li et al., 2024a)	General	CC BY-NC-4.0	36	1	✔	✘	✘
GAOKAO-Bench (Zhang et al., 2023b)	General	Apache-2.0	53	2	✔	✘	✘
XiezhiBenchmark (Gu et al., 2024)	General	CC BY-NC-SA-4.0	85	1	✔	✘	✔
LLMEVAL-2 (Zhang et al., 2023a)	General	-	11	1	✔	✘	✘
ACLUE (Zhang and Li, 2023)	Classical Chinese	CC BY-NC-4.0	1,805	6	✔	✔	✘
WYWEB (Zhou et al., 2023)	Classical Chinese	-	2,500	3	✔	✔	✘
WenMind (Cao et al., 2024a)	Classical Chinese	CC BY-NC-SA-4.0	1,845	16	✔	✔	✔
CCPM (Li et al., 2021)	Ancient Chinese Poetry	-	2,720	1	✔	✘	✘
THU-FSPC (Chen et al., 2019)	Ancient Chinese Poetry	-	5,000	1	✔	✘	✘
ACP-Eval (Ours)	Ancient Chinese Poetry	CC BY-NC-SA-4.0	**7,050**	**24**	✔	✔	✔

Table 13: Comparison of existing evaluation datasets. "HG" indicates Human Generated, "CI" indicates Collection and Improvement of existing datasets, and "MC" indicates Model Constructed. "Scale" for all evaluation datasets refers to the scale of evaluation data related to ancient Chinese poetry.

A.3 Positive and Negative Sample Pairs

The steps for generating positive and negative sample pairs are as follows:

(1) Prompt Design: Taking the task "Title to Author" as an example, we use the prompt "Task: Inquire about the author based on the title of ancient poetry. You are an expert in question generation, please generate several question templates for this task. For example: (a) Who is the author of the ancient poem 'xxx'? (b) Please tell me the author of 'xxx'." ERNIE-4.0 (Baidu, 2023) generates question templates, which are manually filtered. We then replace the placeholder ('xxx') with the title of the poem, creating positive "Q-Q (Question)" pairs for training (for tasks where templates don't apply, we directly guide ERNIE-4.0 to rewrite them differently).

(2) Positive and Negative Pair Generation: Negative "Q-Q" pairs form between different tasks.

Additionally, we also randomly generate "Q-A (Answer)" positive and negative pairs.

(3) Training of the Embedding Model and Ranking Model: The semantic comparison between these positive and negative pairs helps the embedding model better understand the semantic structure of user queries, improving its ability to recall relevant knowledge blocks while maintaining efficient retrieval speed. Unlike the embedding model, the ranking model processes both positive and negative pairs during fine-tuning, enhancing semantic interaction between sentences and improving similarity perception in the field of ancient Chinese poetry, leading to higher retrieval accuracy.

B Evaluation Metrics

B.1 Evaluation Metrics for RAG Used in This Paper

B.1.1 Response Accuracy (RA)

Decompose the answer into several scoring points and determine whether each scoring point is addressed in the LLM's response.

Assume there are *N* questions, each with P_i possible scoring points. Determine how many of these scoring points are present in the responses from LLMs, denoted as S_i (the number of valid scoring points). The Response Accuracy (RA) is calculated using the following formula:

$$RA = \frac{\sum_{i=1}^{N} S_i}{\sum_{i=1}^{N} P_i} \quad (5)$$

This metric relies on task definitions and structures for answer decomposition. Scoring criteria are task-specific, with examples and requirements provided for each. For instance, in the "Content to Title" task, the score is based on accurately identifying the poem's title. In the "Poetry Appreciation" task, scores are divided across dimensions such as emotion and imagery.

B.1.2 Response Continuity (RC)

Based on the responses of LLMs, Response Continuity is assessed from four aspects: whether the response is coherent and fluid, whether there are grammatical errors, whether there is any sentence truncation, and whether there is content repetition.

Assume there are *N* questions, each question has four evaluation points scored as S_{i1}, S_{i2}, S_{i3} and S_{i4}, each scoring either 0 or 0.25. The Response Continuity (RC) is calculated as the average of the total evaluation scores across all questions, using the following formula:

$$RC = \frac{\sum_{i=1}^{N}(S_{i1} + S_{i2} + S_{i3} + S_{i4})}{N} \quad (6)$$

B.1.3 Response Relevance (RR)

Decompose the responses of LLMs into several key points and determine whether each point is relevant to the question.

Assume there are *N* questions, and the LLM's response can be broken down into P_i points for each question. Determine how many of these points are associated with the question, denoted as S_i (the number of effective points). The Response Relevance (RR) is calculated using the following formula:

$$RR = \frac{\sum_{i=1}^{N} S_i}{\sum_{i=1}^{N} P_i} \quad (7)$$

B.1.4 Context Information Volume (CIV)

The context is decomposed into several key points. Based on the query and answer, each key point is evaluated to determine whether it contributes to the generation of the answer. If it is beneficial, it is classified as a valid key point. This metric calculates the proportion of valid information within the context, which represents the accuracy of the context.

Assume there are *N* questions. For each question, the context is decomposed into P_i key points. The number of key points that contribute to the generation of the answer is denoted as S_i (the number of beneficial key points). The Context Information Volume (CIV) is calculated using the following formula:

$$CIV = \frac{\sum_{i=1}^{N} S_i}{\sum_{i=1}^{N} P_i} \quad (8)$$

B.1.5 Context Match Score (CMS)

The answer is decomposed into several scoring points. Each scoring point is assessed to determine whether the context contains relevant information related to it. This metric calculates the proportion of scoring points in the answer that can be matched with the context, which represents the recall rate of the context.

Assume there are *N* questions. For each question, the answer is decomposed into P_i scoring points. The number of scoring points in the context that contain relevant information is denoted as S_i. The Context Match Score (CMS) is calculated using the following formula:

$$CMS = \frac{\sum_{i=1}^{N} S_i}{\sum_{i=1}^{N} P_i} \tag{9}$$

B.1.6 Context Topic Relevance (CTR)

The relevance of the themes involved between the query and the context is assessed. This metric employs a scoring system: 0 points for completely unrelated themes, 1 point for themes that are partially related, 2 points for themes that are largely related, and 3 points for themes that are closely related.

Assume there are N questions. For each question, the relevance score of the themes between the query and the context is denoted as S_i (with a range from 0 to 3). The Context Topic Relevance (CTR) is calculated using the following formula:

$$CTR = \frac{\sum_{i=1}^{N} S_i}{3N} \tag{10}$$

B.2 Comparison with Other RAG Evaluation Metrics

Table 14 compares the RAG metrics proposed in this paper with other RAG metrics. It is observed that existing RAG metrics primarily use LLMs for evaluation, focusing mainly on Response Relevance, Context Match Score, and Context Topic Relevancee. In contrast, this paper evaluates the response from three perspectives (Accuracy, Continuity, and Relevance) and the context from three perspectives (Information Volume, Match Score, and Topic Relevance) based on the relationships among the four elements (Query, Context, Response and Answer). This approach provides a more comprehensive evaluation compared to other RAG evaluation systems.

B.3 Scoring Model

B.3.1 Scoring Prompt

Figure 12 and 13 present the prompts used for evaluating six metrics with the scoring model.

B.3.2 Fine-tuning the Scoring Model

We choose to fine-tune the Qwen1.5-7B base model (Bai et al., 2023) using scoring instructions. The construction process of these scoring instructions is as follows: (1) We refer to 24 types of tasks in ancient Chinese poetry and additionally create 500 Q&A pairs, which do not overlap with ACP-Eval. (2) We use these 500 Q&A pairs to evaluate the LLaMa2-Chinese-7B-Chat (FlagAlpha, 2024), ChatGLM3-6B (Du et al., 2022), Spark-3.5 (Iflytek, 2023), and Yi-9B-Chat model (Young et al., 2024), resulting in a total of 2,000 data points. (3) We evaluate the 2,000 Q&A data points using the ERNIE-3.5 model (Baidu, 2023) across 6 metrics, ultimately generating 12,000 scoring instructions for fine-tuning the scoring model. In this way, the fine-tuned model can learn the key points and formatting requirements of scoring instructions.

B.4 Human Consistency Analysis

To ensure the reliability of the scoring model's results, we conduct a Human Consistency Analysis to verify the correctness and alignment with human preferences of the scores. First, we randomly select 240 questions from ACP-Eval according to task proportions and obtain the corresponding contexts and answers from GPT-4 (OpenAI et al., 2024), ERNIE-4 (Baidu, 2023), LangChain-ChatChat (Liu et al., 2024a), and Self-RAG (Asai et al., 2023), resulting in a total of 960 entries. Next, the scoring model evaluates these 960 results based on 6 metrics, yielding 4,320 entries. Finally, following the principle of "majority rules", three human volunteers assess the reasonableness of the 4,320 scoring results (the volunteers are graduate students from the fields of electronic information and linguistics). If two or more volunteers consider a score to be reasonable, it is deemed correct; otherwise, it is considered incorrect.

We define "reasonable" in scoring in three dimensions: **(1) Consistency:** Each scoring point must align with key aspects of the query and match human expectations. **(2) Logic:** The scoring rationale must clearly explain the basis for the score and reasons for deductions, without logical flaws. **(3) Coverage:** The scoring process should address all core elements of the query without omitting critical points.

The final results, shown in Table 15, indicate that the consistency across the 6 metrics exceeds 86.00%, with an overall human consistency of 92.34%, demonstrating that the scoring model's results are quite reliable.

Name	Evaluation Method	Metrics					
		RA	RC	RR	CIV	CMS	CTR
TruLens-Eval (Truera, 2023)	LLMs	✘	✘	✔	✘	✔	✔
RAGAs (Es et al., 2024)	LLMs + Cosine Similarity	✘	✘	✔	✘	✔	✔
ARES (Saad-Falcon et al., 2024)	LLMs + Classifier	✘	✘	✔	✘	✔	✔
Our Metrics	LLMs	✔	✔	✔	✔	✔	✔

Table 14: Comparison of different RAG evaluation systems.

Model	RA	RC	RR	CIV	CMS	CTR	Overall
GPT-4	87.92	97.92	95.42	-	-	-	-
ERNIE-4	88.75	98.33	95.42	-	-	-	-
LangChain-ChatChat	87.08	97.92	92.50	86.67	85.42	93.75	-
Self-RAG	90.42	99.17	94.58	87.08	87.92	95.83	-
Overall	88.54	98.33	94.48	86.88	86.67	94.79	**92.34**

Table 15: Results of human consistency analysis.

C Experiments

C.1 Specific Task Evaluation Results

Table 17 through 28 present the specific values of six metrics for all experimental comparison methods across 24 distinct tasks.

C.2 Evaluation Results for Different Dimensions and Instruction Categories.

Table 29 to 34 present the results of six metrics for various methods across different evaluation dimensions and instruction categories. “Open QA” refers to open-domain question answering, where questions typically do not have standard answers; “Closed QA” refers to closed-domain question answering, where questions usually have fixed standard answers. From the response accuracy metrics shown in Table 29, the following conclusions can be drawn:

(1) In terms of dimensions, most methods perform significantly better in the “comprehension” dimension compared to the “knowledge” dimension. This indicates that LLMs have strong comprehension and application abilities, excelling in tasks such as translation and appreciation. However, their performance in knowledge-based question answering is poorer due to limitations in training data and the model’s catastrophic forgetting problem.

(2) Compared to Qwen1.5-7B, ACP-RAG improves performance by 11.5% and 69.0% in the “comprehension” and “knowledge” dimensions, respectively. The introduction of the ACP-RAG retrieval framework significantly enhances the model’s performance in both dimensions, especially in the “knowledge” dimension. ACP-RAG partially alleviates the model’s hallucination problem.

(3) In terms of instruction categories, most methods perform weaker in the “Closed QA” category, as “Closed QA” typically involves pure knowledge tasks with fixed answers.

(4) The introduction of ACP-RAG raises the performance across different instruction categories to above 89.0%, providing a noticeable improvement and making it more suitable for the field of ancient Chinese poetry compared to other RAG methods.

C.3 Comparison of Responses from Different Methods

Figure 14 to 17 present response examples from various methods for the “Title and Author to Content” and “The Origin of Idiom” tasks.

C.4 Evaluation Results on Other Datasets

We select nine open-source evaluation datasets and randomly choose 421 questions related to ancient poetry to construct the evaluation dataset ACP-Others. We evaluate all methods using the ACP-Others dataset, and the results for the “Response Accuracy” metric are presented in Table 35. ACP-RAG continues to demonstrate good performance on other open-source evaluation datasets, achieving an overall score of 79.2%, which is comparable to ERNIE-4.0 (Baidu, 2023).

C.5 Prompt Engineering

Figure 18 illustrates the prompt engineering strategy we employ when integrating context and questions for input into the generation model. Effective prompts enable LLMs to learn context more effectively, leading to higher-quality responses.

C.6 Evaluation Results across Different Dynasties

We select Task 3 (Content to Dynasty) and Task 4 (Content to Three Elements) to evaluate the system's performance across different dynasties in ancient Chinese poetry. The results are shown in Table 16. It can be seen that RA and CMS tend to be higher in dynasties with fewer poems (e.g., Sui and Jin), where it is easier to retrieve the correct content, though CIV is lower, making retrieval more susceptible to influence from data sources of other dynasties. Conversely, the opposite pattern is observed in dynasties with larger volumes of poetry.

Dynasty	RA	CIV	CMS
Sui	**98.6**	20.0	98.7
Tang	94.9	50.0	91.7
Song	94.1	49.8	95.6
Jin	97.1	37.2	**99.0**
Yuan	96.7	61.3	97.7
Ming	94.6	**67.5**	95.8
Qing	95.6	67.1	95.7

Table 16: Performance across different dynasties. "RA" represents Response Accuracy, "CIV" represents Context Information Volume, and "CMS" represents Context Match Score.

D Statement of Responsibility

The licenses for ACP-Corpus, ACP-QA, and ACP-Eval are CC-BY-NC-SA-4.0, and they strictly adhere to the agreements of the original data sources.

The ACP-Corpus, ACP-QA, and ACP-Eval datasets involve ancient Chinese poetry, which may encompass traditional views, themes of war, life and death, and sacrificial practices, as well as potential biases. Therefore, we emphasize that these datasets are intended solely for academic research, aimed at analyzing and examining the knowledge and cultural values inherent in ancient poetry. Researchers utilizing these datasets are required to adhere to relevant ethical standards and refrain from using the datasets for inappropriate purposes or commercial gain.

E Ethical Considerations

When utilizing datasets of ancient Chinese poetry, we need to be aware of potential historical biases and stereotypes they may contain, avoiding the reinforcement of outdated social notions by the model. Simultaneously, we must respect the cultural connotations of ancient Chinese poetry, avoiding improper interpretations. We strictly adhere to the original copyrights of the data, with these datasets and model used solely for academic research purposes, aiming to promote the inheritance and innovation of cultural heritage. Commercial use or applications that violate ethical principles are strictly prohibited.

Metric: Response Accuracy (RA)

你是一个严谨、遵守规则的改卷老师。请根据上面提供的问题、参考答案和学生作答内容，完成以下任务：
（1）作为改卷老师，你需要将参考答案分解为若干个得分点。
（2）根据每一个得分点，判断学生作答内容中是否有涉及得分点的内容。
请给出参考答案中你设置的得分点个数，学生作答内容中符合的得分点个数，以及相关理由。
请注意：学生作答内容中符合的得分点个数一定≤参考答案中你设置的得分点个数（至少为1）。

输出格式要求如下，请不要输出其他额外内容：["参考答案中你设置的得分点个数", "学生作答内容中符合的得分点个数", "理由"]
例如：["4", "3", "理由..."]

You are a meticulous, rule-abiding grader. Based on the provided questions, reference answers, and student responses, please complete the following tasks:
(1) As a grader, you need to break down the reference answer into several scoring points.
(2) For each scoring point, determine whether the student's response includes content that addresses the scoring point.
Please provide the number of scoring points you set in the reference answer, the number of scoring points addressed in the student's response, and the relevant reasons.
Note: The number of scoring points addressed in the student's response must be ≤ the number of scoring points you set in the reference answer (at least 1).

The output format should be as follows, and please do not include any additional content: ["Number of scoring points set in the reference answer", "Number of scoring points addressed in the student's response", "Reason"]
For example: ["4", "3", "Reason..."]

Metric: Response Continuity (RC)

你是一个严谨、遵守规则的判断专家。请根据上面提供的问题和回答，仅对"回答"进行判断：
回答是否通顺流畅？若通顺流畅输出"是"，否则输出"否"；
回答是否存在语法错误？若存在语法错误输出"是"，否则输出"否"；
回答是否存在句子截断？若存在句子截断输出"是"，否则输出"否"；
回答是否存在内容重复？若存在内容重复输出"是"，否则输出"否"；
规则：请按顺序给出判断并给出理由。输出格式要求如下，请不要输出其他额外内容：
["是或否", "是或否", "是或否", "是或否", "你给的理由"]

You are a meticulous, rule-abiding evaluator. Based on the provided questions and answers, please evaluate the "answers" only:
Is the answer smooth and coherent? If yes, output "Yes"; otherwise, output "No".
Are there any grammatical errors in the answer? If yes, output "Yes"; otherwise, output "No".
Is there any sentence truncation in the answer? If yes, output "Yes"; otherwise, output "No".
Is there any content repetition in the answer? If yes, output "Yes"; otherwise, output "No".
Rule: Please provide your evaluations in order and give reasons. The output format should be as follows, and please do not include any additional content:
["Yes or No", "Yes or No", "Yes or No", "Yes or No", "Your reasons"]

Metric: Response Relevance (RR)

你是一个严谨、遵守规则的改卷老师。请根据上面提供的问题和文本，完成以下任务：
（1）作为改卷老师，你需要将上面的文本分解为若干个要点。
（2）请判断你分解的每一个要点所涉及的主题是否与问题的主题相关。
请给出文本中你设置的要点个数，与问题的主题相关的要点个数，以及相关理由。
请注意：与问题的主题相关的要点个数一定要≤文本中你设置的要点个数（至少为1）。

输出格式要求如下，请不要输出其他额外内容：["文本中设置的要点个数", "与问题的主题相关的要点个数", "理由"]
例如：["4", "3", "你给的理由"]

You are a meticulous, rule-abiding grader. Based on the provided questions and text, please complete the following tasks:
(1) As a grader, you need to break down the above text into several key points.
(2) Determine whether the topic of each key point you identified is relevant to the topic of the question.
Please provide the number of key points you set in the text, the number of key points relevant to the topic of the question, and the relevant reasons.
Note: The number of key points relevant to the topic of the question must be ≤ the number of key points you set in the text (at least 1).

The output format should be as follows, and please do not include any additional content: ["Number of key points set in the text", "Number of key points relevant to the topic of the question", "Reasons"]
For example: ["4", "3", "Your reasons"]

Figure 12: Scoring prompts for the Response Accuracy, Response Continuity, and Response Relevance metrics.

Metric: Context Information Volume (CIV)

你是一个严谨、遵守规则的改卷老师。请根据上面提供的问题、标准答案和文本，完成以下任务：
（1）作为改卷老师，你需要将上面的文本分解为若干个内容要点。
（2）请参考问题和相应的标准答案，判断文本中你分解的每一个内容要点是否有助于答案的生成。
请给出文本中你设置的内容要点个数，其中有助于答案生成的内容要点个数，以及相关理由。
请注意：有助于答案生成的内容要点个数一定要≤文本中你设置的内容要点个数（至少为1）。

输出格式要求如下，只按照输出格式输出内容，请不要输出其他额外内容：[“文本中设置的内容要点个数”, “有助于答案生成的内容要点个数”, “理由”]
例如：[“4”, “3”, “你给的理由”]

You are a meticulous, rule-abiding grader. Based on the provided questions, model answers, and text, please complete the following tasks:
(1) As a grader, you need to break down the above text into several content points.
(2) Referring to the questions and corresponding model answers, determine whether each content point you identified in the text contributes to the generation of the answer.
Please provide the number of content points you set in the text, the number of content points contributing to the generation of the answer, and the relevant reasons.
Note: The number of content points contributing to the generation of the answer must be ≤ the number of content points you set in the text (at least 1).

The output format should be as follows, and please do not include any additional content: [“Number of content points set in the text”, “Number of content points contributing to the generation of the answer”, “Reasons”]
For example: [“4”, “3”, “Your reasons”]

Metric: Context Match Score (CMS)

你是一个严谨、遵守规则的改卷老师。请根据上面提供的文本A和文本B，完成以下任务：
（1）请将文本B看作参考答案，作为改卷老师的你需要将文本B分解为若干个得分点。
（2）根据每一个得分点，看文本A中是否具有相关内容，只要是有就可以，请忽略其他无关内容。
请给出文本B中你设置的得分点个数，文本A中具有的得分点个数，以及相关理由。
请注意：文本A中符合的得分点个数一定≤文本B中你设置的得分点个数（至少为1）。

输出格式要求如下，请不要输出其他额外内容：[“文本B中设置的得分点个数”, “文本A中符合的得分点个数”, “理由”]
例如：[“4”, “3”, “你给的理由”]

You are a meticulous, rule-abiding grader. Based on the provided Text A and Text B, please complete the following tasks:
(1) Consider Text B as the reference answer. As a grader, you need to break down Text B into several scoring points.
(2) For each scoring point, check whether Text A contains related content. If it does, count it, and ignore any unrelated content.
Please provide the number of scoring points you set in Text B, the number of scoring points present in Text A, and the relevant reasons.
Note: The number of scoring points present in Text A must be ≤ the number of scoring points you set in Text B (at least 1).

The output format should be as follows, and please do not include any additional content: [“Number of scoring points set in Text B”, “Number of scoring points present in Text A”, “Reasons”]
For example: [“4”, “3”, “Your reasons”]

Metric: Context Topic Relevance (CTR)

你是一个严谨、遵守规则的打分专家。请根据上面提供的问题和文本，对两者之间的“主题关联性”进行打分。
打分细则如下：
（1）“主题关联性”是指“问题”和“文本”之间所涉及的主题是否紧密相关。
（2）0分：主题完全不相关；1分：主题小部分相关；2分：主题大部分相关；3分：主题紧密相关。
请根据打分细则给出分数和给分理由。输出格式要求如下，请不要输出其他额外内容：[“分数”, “给分理由”]
这是给你参考的输出格式：[“3”, “你给的理由”]

You are a meticulous, rule-abiding scoring expert. Based on the provided question and text, please score the “thematic relevance” between them.
Scoring criteria are as follows:
(1) “Thematic relevance” refers to how closely the themes of the “question” and the “text” are related.
(2) 0 points: The themes are completely unrelated; 1 point: The themes are partially related; 2 points: The themes are mostly related; 3 points: The themes are closely related.
Please provide the score and the reason for the score based on the scoring criteria. The output format should be as follows, and please do not include any additional content: [“Score”, “Reason for the score”]
For example: [“3”, “Your reasons”]

Figure 13: Scoring prompts for the Context Information Volume, Context Match Score, and Context Topic Relevance metrics.

Method	Overall	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
						General LLMs without Retrieval							
LLaMA3-Chinese-8B	38.6	1.2	0.0	0.7	0.1	75.1	69.6	50.8	0.0	63.8	0.5	6.4	26.3
Baichuan2-7B	41.9	3.4	0.6	7.6	0.7	81.3	73.3	64.4	1.0	72.1	0.7	2.2	21.3
GPT-4	45.8	2.0	0.4	4.8	0.0	82.8	80.5	68.9	0.0	65.2	0.0	3.0	34.8
Qwen1.5-7B	49.2	1.6	0.8	7.4	0.8	87.2	83.3	78.5	0.5	75.3	0.3	3.5	30.7
						LLMs for Classical Chinese without Retrieval							
Bloom-7B-Chunhua	29.4	0.1	0.0	2.4	0.4	69.2	52.9	38.5	0.2	55.8	0.0	2.7	8.3
Xunzi1.5	35.5	8.4	1.0	13.3	0.6	81.8	36.3	54.7	0.7	63.7	0.8	3.7	14.3
Qwen1.5-7B-SFT	51.3	6.2	7.0	50.0	10.5	84.8	60.5	85.4	23.2	84.0	18.0	1.7	24.6
						Industry RAG							
Perplexity.ai	52.2	42.9	55.2	29.9	39.6	82.8	77.0	65.2	25.2	80	18	18.7	40.1
Kimi	75.4	80.7	87.4	63.8	80.7	89.4	81.4	86.7	76.5	78.3	75	**95.5**	79.6
ERNIE-4.0	76.8	77.6	88.6	65.3	9.8	89.7	87.0	88.5	93.8	77	74	86.5	72.1
						Reproducible RAG Methods							
SAIL	36.0	50.2	57.3	71.4	59.7	54.8	28.0	25.0	28.2	66.0	19.2	9.5	1.4
LLaMAIndex-RAG	49.8	2.4	0.6	24.7	5.2	89.3	80.2	80.5	0.7	76.7	1.3	2.4	12.0
LangChain-ChatChat	56.1	12.7	17.1	40.0	10.0	88.4	81.6	84.2	0.0	84.8	0.0	8.5	13.9
Self-RAG	71.5	90.3	97.4	96.9	93.9	88.0	85.8	82.3	73.4	90.9	25.7	15.3	13.4
ACP-RAG (Ours)	89.0	**93.1**	**98.1**	95.9	95.6	90.2	86.4	89.7	85.5	91.0	82.9	92.8	79.0
ACP-RAG + SFT (Ours)	**92.4**	88.6	91.9	**98.5**	**98.8**	**99.7**	**91.5**	**96.9**	**94.6**	**98.3**	**87.0**	94.0	**86.8**

Table 17: Response Accuracy metrics for tasks 1 through 12. Tasks corresponding to "T1" through "T12" are referenced in Table 2. Subsequent tables will not repeat this information.

Method	Overall	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24
						General LLMs without Retrieval							
LLaMA3-Chinese-8B	38.6	2.0	49.6	43.1	26.9	18.8	34.2	34.3	60.8	5.6	45.0	60.3	3.0
Baichuan2-7B	41.9	34.4	25.4	46.7	45.5	31.3	35.4	40.6	61.6	37.8	53.6	66.2	21.2
GPT-4	45.8	21.6	**65.6**	55.0	54.5	26.4	48.6	45.2	72.5	20.2	59.5	78.2	0.7
Qwen1.5-7B	49.2	32.4	53.5	58.9	49.3	38.1	39.5	44.1	75.6	77.0	66.1	66.8	23.1
						LLMs for Classical Chinese without Retrieval							
Bloom-7B-Chunhua	29.4	8.8	19.7	23.5	17.8	10.4	17.2	28.4	45.0	16.3	33.2	24.3	15.0
Xunzi1.5	35.5	1.3	5.1	31.2	26.0	13.0	19.4	45.8	46.7	82.9	55.0	43.5	35.8
Qwen1.5-7B-SFT	51.3	99.0	23.0	29.5	32.2	19.3	60.9	69.2	62.3	97.8	75.3	**89.7**	80.8
						Industry RAG							
Perplexity.ai	52.2	11.4	51.1	51.0	66.3	53.6	61.5	62.1	71.7	62.9	80.4	56.5	21.6
Kimi	75.4	38.1	36.2	61.1	56.1	43.7	88.1	57.1	73.8	80.4	56.2	78.6	29.4
ERNIE-4.0	76.8	72.7	59.1	74.4	76.1	82.1	90.4	69.3	82.1	92.3	**83.3**	77.9	23.7
						Reproducible RAG Methods							
SAIL	36.0	28.6	20.0	8.1	12.4	6.4	10.4	23.6	33.1	26.9	28.0	33.3	45.9
LLaMAIndex-RAG	49.8	23.9	38.2	63.9	73.5	82.1	65.8	43.5	67.5	66.8	60.4	38.7	4.0
LangChain-ChatChat	56.1	17.9	59.6	89.5	70.8	81.6	71.1	20.5	65.9	71.9	63.3	34.6	0.0
Self-RAG	71.5	**100.0**	62.3	39.2	56.9	50.0	32.7	68.2	**87.7**	**98.8**	79.7	77.8	41.7
ACP-RAG (Ours)	89.0	98.5	59.5	89.8	81.0	81.1	96.6	75.9	86.9	93.5	77.6	85.2	**92.0**
ACP-RAG + SFT (Ours)	**92.4**	99.5	30.6	**92.9**	**83.5**	**95.1**	**98.9**	**90.6**	85.6	98.1	75.8	89.5	86.0

Table 18: Response Accuracy metrics for tasks 13 through 24. Tasks corresponding to "T13" through "T24" are referenced in Table 2. Subsequent tables will not repeat this information.

Method	Overall	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
General LLMs without Retrieval													
LLaMA3-Chinese-8B	98.1	93.3	95.3	96.9	97.9	99.7	99.9	100.0	99.7	90.4	97.9	97.4	100.0
Baichuan2-7B	98.3	94.9	96.8	98.6	99.8	99.9	100.0	100.0	99.8	83.8	99.1	96.8	99.8
GPT-4	99.4	98.5	99.4	99.9	99.3	100.0	100.0	100.0	99.6	98.1	100.0	99.1	100.0
Qwen1.5-7B	99.5	98.9	99.4	99.6	99.8	100.0	100.0	100.0	99.4	93.1	100.0	99.8	100.0
LLMs for Classical Chinese without Retrieval													
Bloom-7B-Chunhua	95.8	82.8	88.1	90.1	99.7	97.6	99.5	100.0	97.3	79.0	99.1	92.0	99.5
Xunzi1.5	98.3	98.8	99.3	99.6	100.0	99.8	97.2	99.8	97.4	81.0	98.6	95.1	99.3
Qwen1.5-7B-SFT	98.1	100.0	100.0	100.0	100.0	99.2	99.6	99.6	91.6	73.9	100.0	96.8	99.8
Industry RAG													
Perplexity.ai	98.6	97.3	98.9	99.2	99.4	100.0	100.0	100	91.3	86.5	99.8	98.5	100
Kimi	99.5	100.0	100.0	100.0	100.0	100.0	100.0	100	100	90.5	100	99.5	100
ERNIE-4.0	**99.9**	100.0	100.0	100.0	100.0	100.0	100.0	100	100	98.3	100	100	100
Reproducible RAG Methods													
SAIL	94.9	96.9	96.6	99.0	97.9	93.3	95.2	96.7	91.8	79.9	99.1	95.3	94.8
LLaMAIndex-RAG	99.0	97.9	98.9	100.0	99.6	100.0	99.9	100.0	98.1	93.4	99.8	99.4	99.8
LangChain-ChatChat	99.1	98.3	99.3	100.0	100.0	100.0	100.0	100.0	100.0	87.5	99.2	100.0	100.0
Self-RAG	99.5	100.0	100.0	100.0	100.0	99.8	100.0	100.0	99.6	85.6	100.0	100.0	100.0
ACP-RAG (Ours)	99.4	100.0	99.9	100.0	100.0	100.0	100.0	100.0	99.8	82.8	100.0	100.0	100.0
ACP-RAG + SFT (Ours)	98.1	100.0	99.9	100.0	100.0	97.4	99.9	99.6	88.5	72.4	100.0	99.6	100.0

Table 19: Response Continuity metrics for tasks 1 through 12.

Method	Overall	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24
General LLMs without Retrieval													
LLaMA3-Chinese-8B	98.1	100.0	100.0	100.0	100.0	99.0	100.0	100.0	100.0	91.4	99.8	99.8	98.1
Baichuan2-7B	98.3	100.0	100.0	100.0	100.0	99.0	100.0	100.0	100.0	97.6	99.8	100.0	93.9
GPT-4	99.4	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0	96.0	99.9	100.0	96.9
Qwen1.5-7B	99.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	100.0	100.0	98.3
LLMs for Classical Chinese without Retrieval													
Bloom-7B-Chunhua	95.8	99.9	100.0	99.8	99.8	100.0	99.8	99.8	99.6	93.3	99.3	100.0	94.6
Xunzi1.5	98.3	96.8	99.8	98.5	100.0	99.5	99.3	99.6	99.3	99.8	99.7	100.0	98.2
Qwen1.5-7B-SFT	98.1	100.0	93.0	97.3	99.3	95.5	100.0	100.0	100.0	100.0	99.9	100.0	98.0
Industry RAG													
Perplexity.ai	98.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0	100.0	100.0	100.0
Kimi	99.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0	100.0	100.0	100.0
ERNIE-4.0	**99.9**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Reproducible RAG Methods													
SAIL	94.9	97.8	97.8	95.3	96.3	83.0	96.3	91.4	92.3	88.5	92.8	97.5	96.8
LLaMAIndex-RAG	99.0	100.0	99.9	100.0	100.0	100.0	99.9	100.0	99.9	99.3	99.8	100.0	92.5
LangChain-ChatChat	99.1	100.0	99.2	100.0	100.0	100.0	100.0	100.0	100.0	98.3	100.0	100.0	95.8
Self-RAG	99.5	100.0	97.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.3
ACP-RAG (Ours)	99.4	100.0	99.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.9
ACP-RAG + SFT (Ours)	98.1	100.0	93.4	99.5	99.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.2

Table 20: Response Continuity metrics for tasks 13 through 24.

Method	**Overall**	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**T11**	**T12**
General LLMs without Retrieval													
LLaMA3-Chinese-8B	77.6	46.8	32.6	47.5	83.1	99.3	99.6	85.1	44.3	96.6	39.4	75.5	91.8
Baichuan2-7B	91.6	68.1	75.3	71.9	99.2	99.8	100.0	80.0	81.7	100.0	87.4	87.0	99.8
GPT-4	79.0	69.7	50.1	57.2	87.6	99.9	99.8	85.3	42.6	99.0	41.3	73.6	87.7
Qwen1.5-7B	84.1	70.9	56.9	64.8	92.3	100.0	100.0	85.0	46.1	100.0	48.4	81.6	92.3
LLMs for Classical Chinese without Retrieval													
Bloom-7B-Chunhua	86.2	48.8	59.1	71.9	91.4	95.3	99.9	81.8	71.5	97.6	73.1	88.0	99.5
Xunzi1.5	92.8	94.3	94.2	97.1	99.3	99.7	95.7	90.1	67.9	98.6	71.7	91.8	99.4
Qwen1.5-7B-SFT	<u>94.8</u>	100.0	100.0	100.0	100.0	99.9	96.8	91.6	24.4	99.5	98.8	100.0	95.4
Industry RAG													
Perplexity.ai	94.0	90.8	93.1	88.6	97.4	100.0	100.0	95.0	31.0	99.3	94.0	91.8	89.2
Kimi	90.4	98.4	45.8	67.0	95.6	99.6	100.0	93.3	98.2	100.0	28.6	88.1	96.5
ERNIE-4.0	86.3	94.4	81.3	77.4	92.1	100.0	100.0	87.9	38.0	99.6	56.8	94.7	98.1
Reproducible RAG Methods													
SAIL	83.2	88.9	85.2	90.6	94.2	94.4	92.0	68.9	52.8	89.9	93.1	86.9	80.0
LLaMAIndex-RAG	89.3	78.8	82.9	82.0	94.2	100.0	99.0	87.2	59.1	99.7	88.3	74.8	70.2
LangChain-ChatChat	87.2	73.1	76.8	84.3	94.6	100.0	97.2	88.5	54.1	98.8	71.4	66.7	61.2
Self-RAG	92.7	99.8	99.5	98.2	99.2	99.8	99.9	91.4	80.9	100.0	79.2	74.4	76.2
ACP-RAG (Ours)	**96.9**	99.5	99.5	98.3	99.8	100.0	99.9	89.2	83.4	99.8	95.7	96.0	92.1
ACP-RAG + SFT (Ours)	94.4	99.8	99.3	100.0	100.0	99.5	96.8	90.4	17.4	99.5	99.8	100.0	92.8

Table 21: Response Relevance metrics for tasks 1 through 12.

Method	**Overall**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**	**T21**	**T22**	**T23**	**T24**
General LLMs without Retrieval													
LLaMA3-Chinese-8B	77.6	88.2	99.7	96.5	92.0	97.9	91.3	89.7	93.9	73.1	71.2	95.0	62.6
Baichuan2-7B	91.6	95.1	98.9	98.3	97.9	100.0	90.8	91.2	97.3	89.4	88.6	99.6	73.4
GPT-4	79.0	75.1	98.3	96.2	97.3	93.0	96.5	84.4	91.0	72.2	73.4	98.1	53.9
Qwen1.5-7B	84.1	76.3	99.5	99.5	98.6	98.2	97.1	88.0	97.1	93.6	76.3	96.4	83.8
LLMs for Classical Chinese without Retrieval													
Bloom-7B-Chunhua	86.2	99.5	100.0	96.4	94.9	97.3	92.9	90.1	93.7	68.8	87.4	66.0	45.0
Xunzi1.5	92.8	68.0	87.5	95.4	98.6	100.0	96.7	95.2	97.0	97.7	86.1	100.0	72.6
Qwen1.5-7B-SFT	<u>94.8</u>	100.0	57.0	83.6	96.2	100.0	100.0	91.3	96.8	100.0	96.5	100.0	94.6
Industry RAG													
Perplexity.ai	94.0	100.0	99.5	97.8	99.5	99.1	99.4	92.3	99.4	92.3	91.9	98.2	74.5
Kimi	90.4	80.2	99.7	99.3	100.0	100.0	100.0	85.7	94.0	94.4	98.3	99.2	81.3
ERNIE-4.0	86.3	90.8	100.0	96.3	99.5	100.0	99.6	91.1	91.6	97.4	68.8	96.3	78.8
Reproducible RAG Methods													
SAIL	83.2	81.6	85.9	81.6	80.1	70.2	88.7	52.0	80.6	69.6	68.4	90.7	85.4
LLaMAIndex-RAG	89.3	95.4	96.1	96.0	98.3	99.5	92.6	78.6	97.2	93.4	90.7	97.4	54.9
LangChain-ChatChat	87.2	94.4	91.9	94.7	96.2	98.9	95.5	77.0	97.8	84.1	83.7	97.2	67.4
Self-RAG	92.7	100.0	84.5	90.0	91.1	91.0	76.7	84.2	99.1	100.0	96.1	97.4	82.0
ACP-RAG (Ours)	**96.9**	99.5	85.9	97.9	98.0	99.5	99.3	89.0	98.5	99.3	95.0	95.5	91.5
ACP-RAG + SFT (Ours)	94.4	100.0	56.0	78.2	97.7	99.6	99.8	92.4	96.8	100.0	97.2	100.0	94.9

Table 22: Response Relevance metrics for tasks 13 through 24.

Method	**Overall**	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**T11**	**T12**
Reproducible RAG Methods													
SAIL	21.1	36.1	<u>27.6</u>	11.2	39.5	34.5	37.4	11.2	14.2	29.2	<u>12.6</u>	14.1	14.0
LLaMAIndex-RAG	18.0	2.7	2.2	9.1	6.6	44.1	38.2	17.7	11.4	44.5	3.3	1.2	12.8
LangChain-ChatChat	<u>40.6</u>	11.6	11.5	**28.4**	19.0	**88.9**	84.5	**52.7**	20.7	**87.7**	2.9	<u>15.5</u>	<u>29.7</u>
Self-RAG	32.0	<u>38.6</u>	26.9	16.0	<u>88.8</u>	41.9	**96.2**	19.6	<u>37.6</u>	<u>79.3</u>	5.9	4.4	6.9
ACP-RAG (Ours)	**63.1**	**89.5**	**70.3**	<u>27.7</u>	**96.7**	<u>86.2</u>	<u>91.7</u>	<u>27.3</u>	**85.5**	74.2	**72.1**	**87.3**	**60.3**
ACP-RAG + SFT (Ours)	**63.1**	**89.5**	**70.3**	<u>27.7</u>	**96.7**	<u>86.2</u>	<u>91.7</u>	<u>27.3</u>	**85.5**	74.2	**72.1**	**87.3**	**60.3**

Table 23: Context Information Volume metrics for tasks 1 through 12.

Method	Overall	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24
Reproducible RAG Methods													
SAIL	21.1	11.5	15.6	17.6	16.6	33.2	13.7	12.6	22.7	14.4	15.0	9.0	10.4
LLaMAIndex-RAG	18.0	9.5	13.8	31.5	27.9	79.0	22.3	18.8	37.6	24.6	36.5	**80.3**	11.4
LangChain-ChatChat	40.6	2.1	13.6	74.3	73.3	**98.2**	**66.7**	24.3	51.9	28.1	28.4	15.1	11.4
Self-RAG	32.0	**28.3**	35.5	12.4	14.8	38.6	14.7	13.7	53.7	17.8	23.5	18.2	9.3
ACP-RAG (Ours)	**63.1**	27.2	**37.4**	**83.6**	**84.8**	87.0	37.4	**53.3**	**59.2**	**47.9**	**50.4**	36.8	**28.1**
ACP-RAG + SFT (Ours)	**63.1**	27.2	**37.4**	**83.6**	**84.8**	87.0	37.4	**53.3**	**59.2**	**47.9**	**50.4**	36.8	**28.1**

Table 24: Context Information Volume metrics for tasks 13 through 24.

Method	Overall	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
Reproducible RAG Methods													
SAIL	32.5	39.4	39.0	39.3	38.3	39.4	38.5	31.2	38.3	39.4	28.0	9.5	29.1
LLaMAIndex-RAG	31.0	2.8	0.4	33.7	3.1	69.2	42.8	30.5	0.7	46.7	0.8	0.0	29.9
LangChain-ChatChat	52.6	42.1	13.0	43.6	14.4	96.4	70.4	48.5	3.3	95.2	0.0	7.0	2.0
Self-RAG	69.2	96.9	95.7	**100.0**	**98.1**	**99.9**	**99.3**	63.7	78.2	**100.0**	32.8	5.9	17.3
ACP-RAG (Ours)	**92.3**	**100.0**	**99.4**	**100.0**	97.9	99.8	97.5	**79.7**	**97.0**	**100.0**	**81.8**	**87.0**	**80.5**
ACP-RAG + SFT (Ours)	**92.3**	**100.0**	**99.4**	**100.0**	97.9	99.8	97.5	**79.7**	**97.0**	**100.0**	**81.8**	**87.0**	**80.5**

Table 25: Context Match Score metrics for tasks 1 through 12.

Method	Overall	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24
Reproducible RAG Methods													
SAIL	32.5	37.8	33.0	31.9	29.8	38.7	34.5	25.8	34.4	31.6	25.8	37.7	36.6
LLaMAIndex-RAG	31.0	2.6	18.8	52.6	61.3	94.4	41.6	24.6	38.9	47.7	33.9	16.9	6.9
LangChain-ChatChat	52.6	2.6	15.2	82.7	73.1	89.7	75.6	23.3	37.5	55.9	36.4	0.0	6.5
Self-RAG	69.2	96.5	**88.3**	3.4	28.4	64.9	6.0	56.7	81.7	**87.2**	64.6	77.0	39.1
ACP-RAG (Ours)	**92.3**	**98.0**	82.0	**93.1**	**81.5**	**98.4**	**89.5**	**73.2**	**88.2**	83.9	**80.0**	**96.0**	**93.0**
ACP-RAG + SFT (Ours)	**92.3**	**98.0**	82.0	**93.1**	**81.5**	**98.4**	**89.5**	**73.2**	**88.2**	83.9	**80.0**	**96.0**	**93.0**

Table 26: Context Match Score metrics for tasks 13 through 24.

Method	Overall	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
Reproducible RAG Methods													
SAIL	53.6	61.1	61.0	60.8	61.1	60.8	58.3	52.0	43.0	60.8	44.8	40.8	39.6
LLaMAIndex-RAG	51.1	36.3	24.8	68.1	56.3	59.5	57.1	47.9	53.5	40.2	23.4	33.0	54.0
LangChain-ChatChat	56.7	26.7	31.5	73.3	70.0	82.2	76.7	76.7	15.6	80.0	37.6	25.6	36.7
Self-RAG	82.5	97.2	**99.8**	**99.8**	**99.8**	94.5	**98.7**	83.0	62.1	**99.7**	50.0	25.8	56.0
ACP-RAG (Ours)	**91.4**	**99.8**	**99.8**	99.6	99.7	**98.9**	95.6	**84.6**	**89.5**	99.3	**83.8**	**88.3**	**74.7**
ACP-RAG + SFT (Ours)	**91.4**	**99.8**	**99.8**	99.6	99.7	**98.9**	95.6	**84.6**	**89.5**	99.3	**83.8**	**88.3**	**74.7**

Table 27: Context Topic Relevance metrics for tasks 1 through 12.

Method	Overall	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24
Reproducible RAG Methods													
SAIL	53.6	60.8	58.3	44.6	43.0	47.3	43.4	50.3	54.7	53.4	52.8	58.9	54.7
LLaMAIndex-RAG	51.1	45.2	27.3	57.3	71.3	80.0	**71.1**	62.2	64.6	58.3	82.8	67.3	37.3
LangChain-ChatChat	56.7	28.9	32.2	75.6	80.5	**96.8**	70.0	61.1	69.0	66.7	76.7	69.8	22.2
Self-RAG	82.5	97.5	94.2	57.8	64.3	70.7	67.7	**86.6**	90.4	**96.7**	**93.0**	**93.2**	80.4
ACP-RAG (Ours)	**91.4**	**99.5**	**95.8**	**87.0**	**84.7**	77.3	70.4	85.0	**91.7**	87.3	90.9	93.0	**85.3**
ACP-RAG + SFT (Ours)	**91.4**	**99.5**	**95.8**	**87.0**	**84.7**	77.3	70.4	85.0	**91.7**	87.3	90.9	93.0	**85.3**

Table 28: Context Topic Relevance metrics for tasks 13 through 24.

Method	Overall	Dimension			Instruction Category		
		Comprehension	Knowledge	Translation	Open QA	Closed QA	Classification
General LLMs without Retrieval							
LLaMA3-Chinese-8B	38.6	61.5	15.3	65.0	55.5	17.7	60.3
Baichuan2-7B	41.9	66.8	17.7	71.8	60.1	19.6	66.2
GPT-4	45.8	70.3	21.0	69.8	68.3	23.7	78.2
Qwen1.5-7B	49.2	76.5	21.0	76.5	70.4	25.9	66.8
LLMs for Classical Chinese without Retrieval							
Bloom-7B-Chunhua	29.4	50.8	9.0	59.2	39.1	11.5	24.3
Xunzi1.5	35.5	57.9	14.0	71.2	33.5	18.4	43.5
Qwen1.5-7B-SFT	51.3	75.5	28.7	84.1	52.2	33.9	**89.7**
Industry RAG							
Perplexity.ai	52.2	74.8	42.1	79.2	64.5	42.1	56.5
Kimi	75.4	78.1	73.3	80.2	70.4	75.9	78.6
ERNIE-4.0	76.8	81.3	72.8	79.5	83.9	72.9	77.9
Reproducible RAG Methods							
SAIL	36.0	43.8	28.5	57.6	24.5	29.8	33.3
LLaMAIndex-RAG	49.8	73.4	26.3	77.1	71.7	28.1	38.7
LangChain-ChatChat	56.1	76.3	40.3	80.4	63.9	39.0	34.6
Self-RAG	71.5	84.6	59.1	84.5	75.1	63.8	77.8
ACP-RAG (Ours)	<u>89.0</u>	<u>88.0</u>	<u>90.0</u>	<u>90.6</u>	<u>86.4</u>	<u>89.3</u>	85.2
ACP-RAG + SFT (Ours)	**92.4**	**92.1**	**92.7**	**97.8**	**91.6**	**90.3**	<u>89.5</u>

Table 29: Response Accuracy metrics for different dimensions and instruction categories.

Method	Overall	Dimension			Instruction Category		
		Comprehension	Knowledge	Translation	Open QA	Closed QA	Classification
General LLMs without Retrieval							
LLaMA3-Chinese-8B	98.1	98.3	98.0	97.2	99.9	97.9	99.8
Baichuan2-7B	98.3	97.7	98.7	94.3	99.9	98.7	100.0
GPT-4	99.4	99.1	99.6	98.4	100.0	99.5	100.0
Qwen1.5-7B	<u>99.5</u>	99.2	99.7	97.9	100.0	99.7	100.0
LLMs for Classical Chinese without Retrieval							
Bloom-7B-Chunhua	95.8	96.7	95.2	92.7	99.6	95.5	100.0
Xunzi1.5	98.3	97.7	98.6	95.4	98.4	98.8	100.0
Qwen1.5-7B-SFT	98.1	97.0	98.7	93.7	99.5	98.6	100.0
Industry RAG							
Perplexity.ai	98.6	98.8	98.6	96.6	100.0	98.6	100.0
Kimi	<u>99.5</u>	98.8	99.9	96.8	100.0	99.9	100.0
ERNIE-4.0	**99.9**	99.9	100.0	99.7	100.0	100.0	100.0
Reproducible RAG Methods							
SAIL	94.9	93.6	95.8	92.0	94.5	95.5	97.5
LLaMAIndex-RAG	99.0	98.3	99.4	95.7	99.9	99.4	100.0
LangChain-ChatChat	99.1	97.8	99.8	94.4	100.0	99.6	100.0
Self-RAG	<u>99.5</u>	98.6	100.0	96.8	100.0	99.9	100.0
ACP-RAG (Ours)	99.4	98.5	100.0	96.1	100.0	99.9	100.0
ACP-RAG + SFT (Ours)	98.1	96.8	98.9	92.7	99.9	98.7	100.0

Table 30: Response Continuity metrics for different dimensions and instruction categories.

Method	Overall	Dimension			Instruction Category		
		Comprehension	Knowledge	Translation	Open QA	Closed QA	Classification
General LLMs without Retrieval							
LLaMA3-Chinese-8B	77.6	91.7	66.6	87.5	97.6	68.0	95.0
Baichuan2-7B	91.6	95.2	88.5	94.3	99.5	87.1	99.6
GPT-4	79.0	91.0	70.3	86.9	97.4	71.0	98.1
Qwen1.5-7B	84.1	96.5	75.2	96.8	98.7	76.1	96.4
LLMs for Classical Chinese without Retrieval							
Bloom-7B-Chunhua	86.2	88.9	83.9	86.4	98.4	82.2	66.0
Xunzi1.5	92.8	95.2	91.0	95.0	97.6	90.4	100.0
Qwen1.5-7B-SFT	<u>94.8</u>	96.4	93.6	98.9	97.3	92.5	100.0
Industry RAG							
Perplexity.ai	94.0	97.6	92.3	96.5	97.8	92.3	98.2
Kimi	90.4	97.4	84.4	96.3	98.1	84.1	99.2
ERNIE-4.0	86.3	95.0	81.2	93.8	99.5	81.3	96.3
Reproducible RAG Methods							
SAIL	83.2	85.9	81.2	91.1	86.4	80.1	90.7
LLaMAIndex-RAG	89.3	94.5	84.7	91.6	96.7	85.3	97.4
LangChain-ChatChat	87.2	94.2	82.5	94.5	91.4	82.5	97.2
Self-RAG	92.7	97.1	88.6	96.3	97.0	89.3	97.4
ACP-RAG (Ours)	**96.9**	97.7	96.1	98.6	98.6	95.5	95.5
ACP-RAG + SFT (Ours)	94.4	96.3	92.9	98.8	97.1	91.7	100.0

Table 31: Response Relevance metrics for different dimensions and instruction categories.

Method	Overall	Dimension			Instruction Category		
		Comprehension	Knowledge	Translation	Open QA	Closed QA	Classification
Reproducible RAG Methods							
SAIL	21.1	22.0	20.5	23.9	29.0	19.5	9.0
LLaMAIndex-RAG	18.0	32.2	12.0	30.8	36.5	11.8	**80.3**
LangChain-ChatChat	<u>40.6</u>	<u>44.9</u>	<u>38.2</u>	<u>54.3</u>	<u>71.2</u>	<u>27.7</u>	15.1
Self-RAG	32.0	42.4	26.1	32.3	63.9	26.7	18.2
ACP-RAG (Ours)	**63.1**	**57.8**	**66.8**	**60.9**	**80.2**	**61.0**	<u>36.8</u>
ACP-RAG + SFT (Ours)	**63.1**	**57.8**	**66.8**	**60.9**	**80.2**	**61.0**	<u>36.8</u>

Table 32: Context Information Volume metrics for different dimensions and instruction categories.

Method	Overall	Dimension			Instruction Category		
		Comprehension	Knowledge	Translation	Open QA	Closed QA	Classification
Reproducible RAG Methods							
SAIL	32.5	37.5	28.2	39.1	36.3	27.9	37.7
LLaMAIndex-RAG	31.0	44.4	20.6	53.9	48.6	16.8	16.9
LangChain-ChatChat	52.6	70.7	40.3	89.2	55.8	34.3	0.0
Self-RAG	<u>69.2</u>	<u>91.4</u>	<u>48.7</u>	<u>93.9</u>	<u>79.7</u>	<u>53.3</u>	<u>77.0</u>
ACP-RAG (Ours)	**92.3**	**95.4**	**89.7**	**99.2**	**93.6**	**88.6**	**96.0**
ACP-RAG + SFT (Ours)	**92.3**	**95.4**	**89.7**	**99.2**	**93.6**	**88.6**	**96.0**

Table 33: Context Match Score metrics for different dimensions and instruction categories.

Method	Overall	Dimension			Instruction Category		
		Comprehension	Knowledge	Translation	Open QA	Closed QA	Classification
Reproducible RAG Methods							
SAIL	53.6	56.7	51.7	58.3	53.2	52.5	58.9
LLaMAIndex-RAG	51.1	52.0	50.6	46.7	61.4	49.4	67.3
LangChain-ChatChat	56.7	63.7	52.7	61.5	70.3	51.2	69.8
Self-RAG	<u>82.5</u>	<u>91.6</u>	<u>76.9</u>	<u>89.9</u>	<u>87.4</u>	<u>79.7</u>	**93.2**
ACP-RAG (Ours)	**91.4**	**92.2**	**90.8**	**93.5**	**90.6**	**91.0**	<u>93.0</u>
ACP-RAG + SFT (Ours)	**91.4**	**92.2**	**90.8**	**93.5**	**90.6**	**91.0**	<u>93.0</u>

Table 34: Context Topic Relevance metrics for different dimensions and instruction categories.

Question:

请告诉我《近日》古诗的内容是怎样的？古诗作者为弘历。
Please tell me what the content of the ancient poem “Jin Ri” is like? The author of the poem is Hongli.

Answer:

题目(Title)：近日(Jin Ri)
作者(Author)：弘历(Hongli)
内容(Content)：近日常作阴，云浓风弗作。三朝三寸雪，虽微覆护补。正当地发润，麦苗渥苏土。今晓云复佳，风随晴以午。不知足之心，依然虑楚楚。问已过何由，盼捷意或怃。然讵敢作威，心惟自知苦。

Qwen1.5-7B:

回答：很抱歉，您所提到的《近日》并非由乾隆皇帝（弘历）所作。乾隆帝是中国清代的皇帝，他的诗词作品以《清高宗御制诗集》为代表，但并未收录名为《近日》的诗篇。可能您记错了作者或者这首诗并非出自清代。如果您能提供更详细的信息，我可以帮助您找到相关的古诗。
Response: I’m sorry, but the poem “Jin Ri” you mentioned was not written by Emperor Qianlong (Hongli). Emperor Qianlong was an emperor of the Qing Dynasty in China, and his poetry works are represented by the “Collected Poems of Emperor Gaozong of Qing,” which does not include a poem titled “Jin Ri.” It is possible that you have mistaken the author, or this poem is not from the Qing Dynasty. If you can provide more detailed information, I can help you find the related ancient poem.

Xunzi1.5:

回答/Response: 日日春光好，春光去不还。何人解赏鉴，千古独朱颜。(Content)

ERNIE-4.0:

回答：关于您提到的古诗《近日》及作者弘历，这里需要澄清一点：在文学资料中，直接以“近日”为题的古诗并不常见，且较少有确凿的记录表明乾隆皇帝弘历（即清高宗爱新觉罗·弘历）写过这样一首直接名为《近日》的古诗。由于缺乏具体指向《近日》这首诗的确凿信息，在这里我无法直接给出弘历所写的《近日》古诗的具体内容。
Response: Regarding the ancient poem “Jin Ri” and its author Hongli that you mentioned, it is important to clarify that in literary records, poems directly titled “Jin Ri” are uncommon, and there is little concrete evidence to suggest that Emperor Qianlong (Hongli) of the Qing Dynasty (Aisin Gioro Hongli) wrote a poem directly titled “Jin Ri.” Due to the lack of specific and concrete information pointing to a poem titled “Jin Ri,” I am unable to provide the specific content of an ancient poem titled “Jin Ri” written by Hongli.

Figure 14: Response examples (Part I) for the “Title and Author to Content” task.

Method	Overall	Dataset								
		WenMind	C-Eval	CIF-Bench	CMMLU	GAOKAO-Bench	XiezhiBenchmark	ACLUE	WYWEB	LLMEVAL-2
General LLMs without Retrieval										
LLaMA3-Chinese-8B	45.2	41.2	77.8	68.8	20.0	41.2	75.9	56.6	44.1	0.0
Baichuan2-7B	55.4	53.5	**100.0**	81.3	60.0	58.0	77.1	47.1	47.1	33.3
GPT-4	65.2	63.8	**100.0**	83.3	54.5	71.2	84.6	55.9	58.3	61.1
Qwen1.5-7B	64.9	62.2	<u>88.9</u>	<u>90.6</u>	58.3	81.5	79.6	63.4	54.5	61.1
LLMs for Classical Chinese without Retrieval										
Bloom-7B-Chunhua	40.1	40.5	33.3	74.2	38.5	40.4	63.8	25.6	28.4	0.0
Xunzi1.5	51.6	47.7	77.8	84.6	73.9	63.0	85.1	38.7	43.1	38.9
Qwen1.5-7B-SFT	59.6	54.0	**100.0**	81.8	**95.0**	80.8	**100.0**	71.6	39.0	88.9
Industry RAG										
Perplexity.ai	66.3	67.3	66.7	83.1	65.4	70.9	87.7	45.8	51.5	78.9
Kimi	72.8	77.8	<u>88.9</u>	88.2	60.0	81.4	86.3	40.5	51.1	55.6
ERNIE-4.0	<u>79.0</u>	<u>78.7</u>	**100.0**	**90.9**	<u>85.2</u>	83.0	78.7	<u>83.3</u>	64.6	**100.0**
Reproducible RAG Methods										
SAIL	34.2	37.8	0.0	60.9	21.7	17.3	46.0	29.6	10.8	70.0
LLaMAIndex-RAG	57.3	55.4	77.8	87.5	48.0	70.4	83.0	51.4	42.0	63.2
LangChain-ChatChat	59.8	59.7	66.7	81.3	59.1	75.9	87.0	50.0	41.0	55.6
Self-RAG	75.3	71.7	**100.0**	86.2	85.0	**90.4**	<u>95.7</u>	75.7	**69.2**	**100.0**
ACP-RAG (Ours)	**79.2**	**78.9**	66.7	86.2	70.8	<u>88.7</u>	89.4	**85.9**	<u>65.9</u>	<u>94.7</u>

Table 35: Comparison between ACP-RAG and other methods on other evaluation datasets.

LangChain-ChatChat:

上下文：作于乾隆十七年正月，弘历设曲宴召沈德潜等文臣共制此诗。
Context: Written in the first month of the 17th year of Qianlong's reign, Hongli composed this poem during a banquet where he invited literary ministers such as Shen Deqian to collaborate on its creation.

回答：《近日》这首诗是乾隆十七年正月，乾隆皇帝（弘历）设曲宴，邀请沈德潜等文臣共同创作的。由于诗的内容没有直接提供，所以无法直接告诉你诗的具体内容。
Response: The poem "Jin Ri" was composed in the first month of the 17th year of Qianlong's reign. Emperor Qianlong (Hongli) hosted a banquet and invited literary ministers such as Shen Deqian to collaborate on its creation. Since the content of the poem is not directly provided, I am unable to tell you its specific content.

ACP-RAG (Ours):

上下文：
QA-0：
弘历写的《近日》这首古诗词的完整内容能否告知？近日常作阴，云浓风弗忤。三朝三寸雪，虽微覆护补。正当地发润，麦苗渥苏土。今晓云复佳，风随晴以午。不知足之心，依然虑楚楚。问已过何由，盼捷意或忓。然讵敢作威，心惟自知苦。
Context:
QA-0:
Can you provide the full content of the ancient poem "Jin Ri" written by Hongli? 近日常作阴，云浓风弗忤。三朝三寸雪，虽微覆护补。正当地发润，麦苗渥苏土。今晓云复佳，风随晴以午。不知足之心，依然虑楚楚。问已过何由，盼捷意或忓。然讵敢作威，心惟自知苦。

回答：《近日》这首古诗词的完整内容如下：近日常作阴，云浓风弗忤。三朝三寸雪，虽微覆护补。正当地发润，麦苗渥苏土。今晓云复佳，风随晴以午。不知足之心，依然虑楚楚。问已过何由，盼捷意或忓。然讵敢作威，心惟自知苦。作者是清朝的乾隆皇帝（弘历）。
Response: The full content of the ancient poem "Jin Ri" is as follows: 近日常作阴，云浓风弗忤。三朝三寸雪，虽微覆护补。正当地发润，麦苗渥苏土。今晓云复佳，风随晴以午。不知足之心，依然虑楚楚。问已过何由，盼捷意或忓。然讵敢作威，心惟自知苦。The author is Emperor Qianlong (Hongli) of the Qing Dynasty.

Figure 15: Response examples (Part II) for the "Title and Author to Content" task. Refer to Figure 11 for detailed information on the question.

Question:

成语“闭门觅句”的含义和出处是什么?
What is the meaning and origin of the idiom “闭门觅句”?

Answer:

形容作诗时冥思苦想，静心写作。出处：宋·黄庭坚《病起荆江亭即事》诗之九：“闭门觅句陈无己，对客挥毫秦少游。”
It describes the act of deep contemplation and focused writing when composing poetry. Origin: Song Dynasty, Huang Tingjian’s poem “Sick at Jingjiang Pavilion, Writing About the Situation” number nine: “闭门觅句陈无己，对客挥毫秦少游。”

Qwen1.5-7B:

回答：“闭门觅句”这个成语出自宋代文人杨万里的诗句。它的含义是指在家中、书房里独自构思写作，寻找诗句或文章的灵感，不外出寻求，也不与他人交流。原句是杨万里的一首诗《闲居初夏午睡起二绝句》中的第一句：“梅子留酸软齿牙，芭蕉分绿与窗纱。”
Response: The idiom “闭门觅句” originates from a poem by the Song Dynasty scholar Yang Wanli. It means to stay at home or in a study, composing and seeking inspiration for poetry or prose alone, without going out or communicating with others. The original line is from Yang Wanli’s poem “Idle in Early Summer, Awakening from a Noon Nap” and is the first line: “梅子留酸软齿牙，芭蕉分绿与窗纱。”

Xunzi1.5:

回答: 成语“闭门觅句”出自唐代诗人杜甫的《春夜喜雨》。它的含义是闭门不出，独自寻找诗句，形容诗人独自思考、创作诗歌的状态。
Response: The idiom “闭门觅句” originates from the Tang Dynasty poet Du Fu’s “Spring Night’s Delightful Rain”. It means staying indoors and seeking poetic lines alone, describing a poet’s state of solitary contemplation and creation of poetry.

ERNIE-4.0:

回答：成语“闭门觅句”的含义和出处如下：“闭门觅句”是一个汉语成语，形容作诗时冥思苦想，沉浸在创作的世界中，寻求最佳的词句和表达方式。这个成语体现了诗人或作家在写作时全神贯注、深思熟虑的状态。出处：“闭门觅句”这个成语出自宋代著名诗人黄庭坚的《病起荆江亭即事》一诗，具体而言，是其中的诗之九：“闭门觅句陈无己，对客挥毫秦少游。”
Response: The meaning and origin of the idiom “闭门觅句” are as follows: “闭门觅句” is a Chinese idiom that describes the state of deep contemplation and immersion in the creative process while composing poetry, seeking the best words and expressions. This idiom reflects the poet’s or writer’s focused and thoughtful state during writing. Origin: The idiom “闭门觅句” comes from the famous Song Dynasty poet Huang Tingjian’s poem “Sick at Jingjiang Pavilion, Writing About the Situation”, specifically from poem number nine: “闭门觅句陈无己，对客挥毫秦少游.”

Figure 16: Response examples (Part I) for the “The Origin of Idiom” task.

LangChain-ChatChat:

上下文：成语释义：悯时病俗，常用汉语成语，读音是（mǐn shí bìng sú）。ABCD式组合，联合式结构，悯时病俗意思是：指忧虑时俗。出自宋·曾巩《答李讼书》。一般作定语，指忧虑时俗。成语出处：宋·曾巩《答李讼书》："足下自称有悯时病俗之心，信如是，是足下之有志乎道。"
成语释义：闭门不出，常用汉语成语，读音是（bì mén bù chū）。ABCD式组合，连动式结构，闭门不出意思是：关起门来不外出，指杜绝与外界交往。出自南朝·宋·范晔《后汉书·史弼传》。一般作谓语、宾语，指与外界不交往。
Context: Idiom Explanation: 悯时病俗 (mǐn shí bìng sú) is a commonly used Chinese idiom in ABCD form with a combined structure. It means to worry about the current times and societal customs. Origin: Song Dynasty, Zeng Gong's "Reply to Li Song": "You claim to have a heart that pities the times and is sickened by societal customs; if this is true, it shows your ambition for the Way." It is generally used as an attributive to indicate concern for societal customs.
Idiom Explanation: 闭门不出 (bì mén bù chū) is a commonly used Chinese idiom in ABCD form with a serial verb structure. It means to shut the door and not go out, indicating a refusal to interact with the outside world. Origin: Southern Dynasties, Song, Fan Ye's "Book of Later Han, Biography of Shi Bi": "You claim to have a heart that pities the times and is sickened by societal customs; if this is true, it shows your ambition for the Way." It is generally used as a predicate or object, indicating no interaction with the outside world.

回答："闭门觅句"并不是您提供的已知信息中的成语，它并不是出自杨万里、曾巩或者范晔的作品。"闭门觅句"这个成语一般是指一个人关在房子里寻找诗句或者创作诗词，比喻绞尽脑汁地写文章或创作。它的出处不明确，但与创作文学相关。如果您需要关于"闭门觅句"的详细解释，我可以提供。
Response: "闭门觅句" is not an idiom found in the information you provided, nor does it originate from the works of Yang Wanli, Zeng Gong, or Fan Ye. The idiom "闭门觅句" generally refers to someone staying indoors to seek out poetic lines or create poetry, metaphorically describing the process of racking one's brains to write or create. Its exact origin is unclear, but it is related to literary creation. If you need a detailed explanation of "闭门觅句," I can provide it.

ACP-RAG (Ours):

上下文：
QA-0：能否介绍一下成语"闭门觅句"的起源和它所代表的具体含义？成语释义：闭门觅句，汉语成语，读音是（bì mén mì jù）。ABCD式组合，连动式结构，闭门觅句意思是：形容作诗时冥思苦想。出自宋·黄庭坚《病起荆江亭即事》诗之九。一般作谓语，形容静心写作。成语出处：宋·黄庭坚《病起荆江亭即事》诗之九："闭门觅句陈无己，对客挥毫秦少游。"
QA-1：成语"闭门不出"的含义和出处是什么？成语释义：闭门不出，常用汉语成语，读音是（bì mén bù chū）。ABCD式组合，连动式结构，闭门不出意思是：关起门来不外出，指杜绝与外界交往。出自南朝·宋·范晔《后汉书·史弼传》。一般作谓语、宾语，指与外界不交往。成语出处：南朝·宋·范晔《后汉书·史弼传》："肃竟归田里，称病闭门不出。"
Context:
QA-0: Can you introduce the origin and specific meaning of the idiom "闭门觅句"? Idiom Explanation: "闭门觅句" (bì mén mì jù) is a Chinese idiom in ABCD form with a serial verb structure. It means to deeply ponder and search for poetic lines while composing poetry. Origin: Song Dynasty, Huang Tingjian's poem "Sick at Jingjiang Pavilion, Writing About the Situation," specifically poem number nine: "闭门觅句陈无己，对客挥毫秦少游". It is generally used as a predicate to describe the state of focused writing.
QA-1: What is the meaning and origin of the idiom "闭门不出"? Idiom Explanation: "闭门不出" (bì mén bù chū) is a commonly used Chinese idiom in ABCD form with a serial verb structure. It means to shut the door and not go out, indicating a refusal to interact with the outside world. Origin: Southern Dynasties, Song, Fan Ye's "Book of Later Han, Biography of Shi Bi": "肃竟归田里，称病闭门不出". It is generally used as a predicate or object to indicate no interaction with the outside world.

回答：成语"闭门觅句"的含义是形容作诗时冥思苦想，出处是宋·黄庭坚的《病起荆江亭即事》诗之九："闭门觅句陈无己，对客挥毫秦少游。"
Response: The meaning of the idiom "闭门觅句" is to describe the deep contemplation and intense thinking while composing poetry. Its origin is from Song Dynasty poet Huang Tingjian's poem "Sick at Jingjiang Pavilion, Writing About the Situation", specifically poem number nine: "闭门觅句陈无己，对客挥毫秦少游".

Figure 17: Response examples (Part II) for the "The Origin of Idiom" task. Refer to Figure 13 for detailed information on the question.

Without Prompt Engineering

[Context]
以上内容是提供给你的参考资料。
你要回答的问题是：[Question]

[Context]
The above content is provided for your reference.
The question you need to answer is: [Question]

With Prompt Engineering

[Context]
以上内容是提供给你的参考资料。你是一名古诗词和阅读专家，如果[参考材料]有效，请借鉴[参考材料]的[内容]回答问题，否则根据[自身知识]回答。
你要回答的问题是：[Question]

[Context]
The above content is provided for your reference. You are an expert in ancient poetry and reading. If [reference material] is valid, please use [content from the reference material] to answer the question. Otherwise, answer based on [your own knowledge].
The question you need to answer is: [Question]

Figure 18: The adopted prompt engineering.