

TongGu: Mastering Classical Chinese Understanding with Knowledge-Grounded Large Language Models

Jiahuan Cao^{1,3} Dezhi Peng¹ Peirong Zhang¹ Yongxin Shi¹
Yang Liu^{1,3} Kai Ding^{2,3} Lianwen Jin^{1,3*}

¹South China University of Technology

²Intsig Information Co., Ltd.

³INTSIG-SCUT Joint Lab on Document Analysis and Recognition

jiahuanc@foxmail.com eelwjin@scut.edu.cn

Abstract

Classical Chinese is a gateway to the rich heritage and wisdom of ancient China, yet its complexities pose formidable comprehension barriers for most modern people without specialized knowledge. While Large Language Models (LLMs) have shown remarkable capabilities in Natural Language Processing (NLP), they struggle with Classical Chinese Understanding (CCU), especially in data-demanding and knowledge-intensive tasks. In response to this dilemma, we propose **TongGu** (mean understanding ancient and modern), the first CCU-specific LLM, underpinned by three core contributions. First, we construct a two-stage instruction-tuning dataset ACCN-INS derived from rich classical Chinese corpora, aiming to unlock the full CCU potential of LLMs. Second, we propose Redundancy-Aware Tuning (RAT) to prevent catastrophic forgetting, enabling TongGu to acquire new capabilities while preserving its foundational knowledge. Third, we present a CCU Retrieval-Augmented Generation (CCU-RAG) technique to reduce hallucinations based on knowledge-grounding. Extensive experiments across 24 diverse CCU tasks validate TongGu's superior ability, underscoring the effectiveness of RAT and CCU-RAG. The model and dataset are available at <https://github.com/SCUT-DLVCLab/TongGu-LLM>.

1 Introduction

Classical Chinese is a vital bridge in connecting the present with the wisdom of ancient China, illuminating insights into historical social life and cultural practices. However, the significant linguistic differences between classical and modern Chinese, including vocabulary and syntax, render this invaluable heritage prohibitively challenging for non-experts to understand.

Recent advancements in Large Language Models (LLMs) have demonstrated remarkable capability in addressing various Natural Language Processing (NLP) tasks (Raffel et al., 2020; Zhang et al., 2022; Chung et al., 2024; Chowdhery et al., 2023; Brown et al., 2020; Touvron et al., 2023a,b; OpenAI, 2023), prompting researchers to explore their capabilities in the specialized realm of Classical Chinese Understanding (CCU). However, existing models, including general-purpose and preliminary CCU-specific LLMs (Wptoux, 2023; Xunzi-LLM-of-Chinese-classics, 2024), often struggle with tasks that require large-scale training data or extensive domain knowledge. This predicament primarily stems from two causes: the lack of dedicated instruction-tuning datasets capable of unleashing their full capabilities, and models' innate propensity to generate hallucinations when tackling knowledge-intensive tasks without sufficient factual grounding.

To address these challenges, we present **TongGu**, a pioneering vertical domain LLM as well as the most proficient CCU specialist. We commence with devising an automated pipeline to construct instruction data from classical Chinese texts, resulting in ACCN-INS (short for ancient Chinese instruction), the first publicly available CCU instruction dataset catering to diverse CCU tasks. Subsequently, TongGu undergoes a two-stage instruction tuning, respectively for the optimization of data-hungry and data-efficient tasks. Here, "data-hungry" means that a substantial volume of data is required to attain satisfactory model performance, and "data-efficient" denotes that a small amount of data is sufficient to achieve desired outcomes. It is first fine-tuned on data-hungry tasks using large-scale training data, such as classical to modern Chinese translation, followed by a second stage of fine-tuning on data-efficient tasks such as punctuation with small-scale data. To prevent catastrophic forgetting during the two-stage fine-tuning,

*Corresponding Author.

we propose **Redundancy-Aware Tuning (RAT)**, a novel sparse fine-tuning (*a.k.a.* Parameter-Efficient Fine-Tuning (PEFT)) method that identifies and freezes the most crucial layers for the current task according to layer redundancy. RAT effectively injects new capability to the model while preserving prior learnt knowledge, thereby ensuring the stability and retention of foundational knowledge in TongGu. In addition, we propose an efficient **CCU Retrieval-Augmented Generation (CCU-RAG)** method that significantly mitigates the propensity of hallucinations in knowledge-intensive tasks, further bolstering TongGu’s performance.

To summarize, our contributions are as follows:

- We develop TongGu, a pioneering vertical domain LLM adept at managing a broad spectrum of CCU tasks.
- We design a pipeline for automatically generating instruction data from classical Chinese texts and construct the ACCN-INS dataset, the first classical Chinese instruction data publicly available.
- We propose Redundancy-Aware Tuning (RAT), a sparse fine-tuning method to alleviate catastrophic forgetting in the two-stage fine-tuning.
- To reduce hallucinations in knowledge-intensive classical Chinese tasks for LLMs, we introduce a task-specific efficient Retrieval-Augmented Generation (RAG) method.

2 Related Work

2.1 Large Language Models

Large Language Models (LLMs), such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023a), and Baichuan (Baichuan, 2023), have exhibited unprecedented prowess across numerous NLP tasks. Thanks to dedicated training techniques like instruction tuning, current LLMs can develop not only exceptional general intelligence but also commendable domain-specific specialties. In this context, research into vertical domain LLMs (Roziere et al., 2023; Wu et al., 2023; Yunxiang et al., 2023) emerges as a burgeoning topic and continually fuels endeavors within the community.

2.2 Language Modeling for Classical Chinese Understanding

Early Classical Chinese Understanding (CCU) systems were trained for specific tasks such as translation (Jiang et al., 2023; Chang et al., 2021),

punctuation (Li and Sun, 2009), and named entity recognition (NER) (Yu and Wang, 2020; Han et al., 2018). However, these methods often relied heavily on large amounts of manually annotated data to achieve decent performance. GujiBERT (Wang et al., 2023) utilized large-scale unlabeled classical Chinese corpora for masked pre-training, providing task-specific models with embeddings that encode classical Chinese knowledge. SikuGPT (Chang et al., 2023) leveraged massive classical Chinese corpora for generative pre-training, highlighting the potential of generative pre-training in classical Chinese poetry and prose creation. Bloom-7b-Chunhua (Wptoux, 2023) and Xunzi-Qwen-7B-CHAT (Xunzi-LLM-of-Chinese-classics, 2024), which combined an open-source base model with large-scale classical Chinese corpora, preliminarily investigated the classical Chinese language understanding capabilities of LLMs.

2.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a technique that enhances generation models by incorporating relevant content retrieved from knowledge sources, which has been proven to be effective for mitigating hallucinations in LLMs (Guu et al., 2020; Lewis et al., 2020).

RAG system generally adheres to a workflow encompassing three components: indexing, retrieval, and generation. Recent advancements have focused on enhancing the retrieval component. RETRO (Borgeaud et al., 2022) amalgamates large-scale corpora with pre-trained frozen BERT embeddings. Atlas (Izacard et al., 2023) conducts joint training of a retriever and a sequence-to-sequence model to attain a language model with robust few-shot learning capabilities. Self-RAG (Asai et al., 2024) selectively retrieves knowledge and generates critique tokens to criticize its own output.

Apart from prior works, our proposed CCU-RAG specifically targets knowledge-intensive tasks within classical Chinese question-answering. It centers on endowing TongGu to discern both the timing and content of retrieval, while enhancing the synergistic efficiency between the model and the retrieval system it relies upon.

3 ACCN-INS Dataset

The complexity of manual annotation in classical Chinese Question Answering (QA) tasks requires extensive human expertise, resulting in labor-

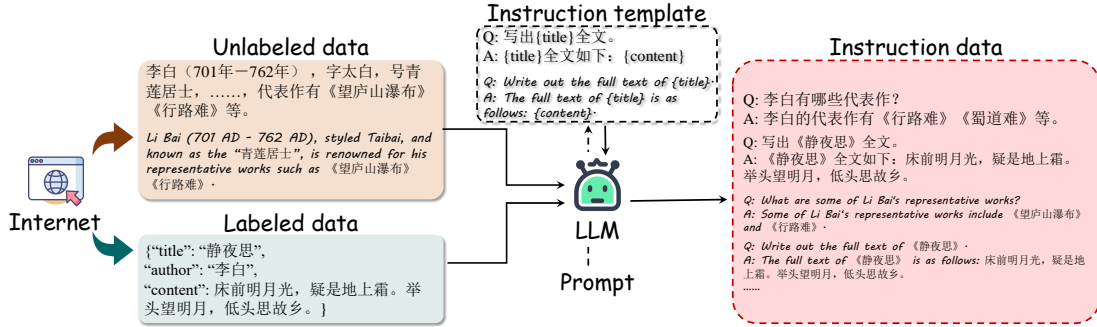


Figure 1: Overview of the CCU instruction data generation pipeline from labeled and unlabeled text.

intensive processes. To alleviate the labor intensity, harnessing LLM for automatic data annotation becomes a natural and efficient solution. Nevertheless, LLM harbors the propensity to inadvertently introduce inaccuracies during the data generation process. To address this issue, we present a semi-automated annotation method that combines classical Chinese corpora using aligned LLMs, thereby producing reliable instructional data for specialized CCU tasks.

Firstly, we collected classical Chinese corpora from multiple public sources, such as Daizhige (Garychowcmu, 2019), textbooks, and examination papers. We then design a pipeline to transform these data into the instructional format, as depicted in Figure 1.

Table 1: Statistics of the generated data.

Statistics	
# instructions	4,020,136
# instructions from labeled data	4,014,355
# data-hungry tasks data	4,000,000
# data-efficient tasks data	14,355
# instructions from unlabeled data	5,781
# data-efficient tasks data	5,781
avg. instruction length	48.59
avg. output length	68.96

Labeled data. Labeled data refers to the data equipped with well-curated labels, such as the dynasty and author of a poem. Owing to its highly structured organization, a large amount of instruction data can be simply synthesized using instruction templates. Specifically, for each task, we first provide 8 instruction examples handcrafted by human experts as in-context examples, prompting the aligned LLM to generate a broader range of diverse instruction templates. Finally, we can populate the structured data into these instruction templates to obtain instruction data in the QA format.

Unlabeled data. Unlabeled data refers to unlabeled text segments, such as introductions to certain poets, where information like era, life experiences, and representative works are intermingled within the same text segment. We adopt a reading comprehension approach, treating the unlabeled text segments as reference materials and requiring the aligned LLM to extract QA pairs from them. Similarly, we use 8 human-written QA pairs as in-context examples.

Employing the proposed data generation pipeline and leveraging ChatGPT (Ouyang et al., 2022) as the aligned model, we initially generated a large corpus of classical Chinese instruction data. To ensure the quality and accuracy of this generated data, we implemented a rigorous cleaning process. This process included several key steps: removing duplicate entries, standardizing punctuation marks to ensure consistency between Chinese and English usage, and carefully reviewing the content for accuracy and appropriateness. After this meticulous verification and refinement process, we obtained a final dataset of 4,020,136 instances of high-quality classical Chinese instruction data, among which 4,014,355 instances originated from structured text and 5,781 from unstructured text. In Table 1, we delineate the quantities of data obtained through various generation methods, coupled with the average lengths of instructions and outputs across the entire instruction dataset. Data-hungry tasks data in ACCN-INS contains 4,000,000 samples of classical-to-modern Chinese translation corpus, and Figure 2 presents the data statistics of various data-efficient tasks data in ACCN-INS. As illustrated in Figure 2 (a), the length distribution indicates that responses are generally longer than queries, with a multitude of responses exceeding 96 characters. This suggests the rich and comprehen-

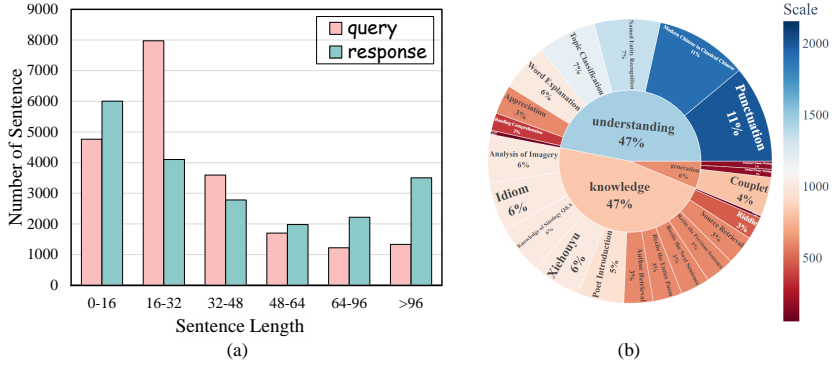


Figure 2: Data statistics of data-efficient tasks data in ACCN-INS dataset. (a) Distribution of sentence length. (b) Sample distribution for each task. Zoom in for a better view.

sive nature of the information in the ACCN-INS dataset, benefiting the model to develop deeper CCU proficiencies. The detailed task types and corresponding sample count of ACCN-INS are illustrated in Figure 2 (b), demonstrating the diversity and comprehensiveness of this dataset. Detailed examples of each task are included in Appendix A.

4 TongGu

TongGu is a generalist LLM specifically designed for Classical Chinese Understanding (CCU), whose capability is built based on three core steps as shown in Figure 3. First, we perform incremental pre-training on TongGu on a mixed corpus of 4.6 billion tokens consisting of classical and modern Chinese to enrich its CCU knowledge. Second, TongGu is fine-tuned on millions of instruction data with the proposed PEFT method Redundancy-Aware Tuning (RAT), which not only enhances multi-task understanding proficiency but also ensures highly efficient instruction tuning. Thirdly, we introduce CCU-RAG, a task-specific Retrieval-augmented generation (RAG) mechanism to alleviate hallucinations in knowledge-intensive tasks. Through these three steps, TongGu demonstrates its ability to handle 24 diverse CCU tasks effectively, making it a powerful tool for understanding classical Chinese texts.

4.1 Incremental Pre-training

For the incremental pre-training of TongGu, we curate hybrid incremental pre-training data consisting of classical Chinese and modern Chinese texts with 2.41 billion tokens in total (with tokenizer from Baichuan2-7B-Base). Table 2 illustrates the various data sources and their respective size and sampling proportions. **Classical Chinese:** The classi-

Table 2: Incremental pre-training data. For each subset, we list the sampling proportion, disk size, and number of epochs during the training process.

Dataset	Sampling prop.	Disk size (GB)	Tokens (B)	Epochs
<i>Classical Chinese</i>				
Series	3.16%	0.27	0.078	1
Buddhism	4.92%	0.42	0.129	2
Confucianism	4.22%	0.36	0.112	2
Medicine	1.52%	0.13	0.042	1
History	17.56%	1.5	0.453	2
Philosophy	1.17%	0.1	0.032	1
Changes	0.94%	0.08	0.026	1
Poetry	8.20%	0.7	0.222	2
Literature	12.88%	1.1	0.315	2
Taoism	0.94%	0.08	0.026	1
Art	4.68%	0.4	0.014	1
<i>Modern Chinese</i>				
wiki-zh	39.9%	3.4	0.968	2
Total	-	8.54	2.41	-

cal Chinese text is primarily sourced from Daizhige (Garychowcmu, 2019) and web-crawled compilations, spanning diverse domains including history, poetry, medicine, and Buddhist studies, etc. We employ a data cleaning pipeline inspired by RedPajama (Together Computer, 2023) to perform text format standardization and document-level deduplication. **Modern Chinese:** We utilize the wiki-zh corpus curated by MNBVC (MOP-LIWU Community and MNBVC Team, 2023) and further perform line-level deduplication.

We utilize the Baichuan2-7B-Base (Baichuan, 2023) as a foundational model and performed incremental pre-training based on the curated mixed data. Following the standard language modeling paradigm outlined in GPT (Radford et al., 2018), we train the model to predict the next token based on the context provided by the previous tokens. As a result, we develop a classical Chinese base model, *TongGu-7B-Base*, poised to serve as a potent foun-

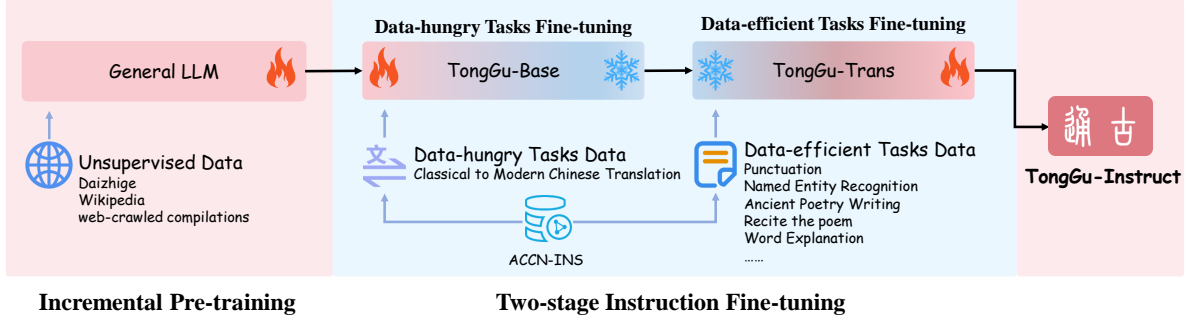


Figure 3: Overview of the training pipeline.

dition for subsequent fine-tuning. More training details such as information on training duration and hardware specifications are provided in Appendix B, Table 9.

4.2 Two-stage Instruction Fine-tuning

Different CCU tasks can be categorized as data-hungry and data-efficient based on their data requirements. The former demands an insatiable feast of data to attain satisfactory performance, epitomized by the Translation task between classical and modern Chinese. The latter, however, is capable of achieving satisfactory performance with modest data provisions, such as punctuation restoration or topic classification. To address both data-hungry and data-efficient task requirements, we conduct a two-stage fine-tuning procedure that first fine-tune TongGu on the data-hungry translation task with a large amount of data and then fine-tune it on data-efficient tasks such as punctuation and topic classification with a smaller scale of data. Through the progressive fine-tuning, the model can effectively capitalize on large-scale data for the primary translation task, while enabling efficient transfer learning and specialization for multiple tasks with limited data, thus fostering comprehensive CCU task proficiency.

Despite the advantage in terms of general ability fostering of the two-stage fine-tuning, this scheme potentially confronts with the catastrophic forgetting issue. To mitigate this issue, we propose a novel PEFT method termed Redundancy-Aware Tuning (RAT). Recent study (Gromov et al., 2024) have revealed that certain layers in LLMs are highly redundant, suggesting they can be removed without significantly impacting the performance of downstream tasks. Building upon this inspiration, RAT identifies and preserves these redundant layers while freezing the others during the training of new

tasks. By selectively updating only the redundant layers deemed non-essential for the erstwhile tasks, this approach effectively retain the acquired knowledge, which, therefore, mitigates catastrophic forgetting, while enabling efficient adaptation to new tasks.

Algorithm 1 summarizes the procedure of RAT. Initially, we randomly select a portion of training data as a calibration set to extract and monitor the model’s internal dynamics. Subsequently, we collect the hidden state representations from each model layer during inference, and calculate the cosine similarity between I/O hidden states in tandem. The cosine similarity between I/O hidden states for the i^{th} layer is calculated as:

$$\text{CoS}_i = \frac{1}{L} \sum_{t=1}^L \frac{H_{i,t} \cdot H_{i+1,t}}{\|H_{i,t}\|_2 \|H_{i+1,t}\|_2} \quad (1)$$

where $H_{i,t}$ represents the hidden state vector at timestep t for layer i , $\|\cdot\|_2$ denotes the L_2 normalization and L represents the sample sentence length. We calculate the mean of the cosine similarity on all samples in the calibration set.

Finally, we freeze the layers exhibiting lower similarity scores. Due to the tendency of deeper layers to harbor a greater degree of redundancy, we have implemented a grouping and ranking strategy to avoid potential impairment to the model’s learning capacity deriving from solely fine-tuning the deeper layers. The layers of TongGu are partitioned into N groups according to their depth, from the shallowest to the deepest. Within each group, we selectively subject the layer exhibiting the highest redundancy to fine-tuning, while the remaining layers are kept frozen.

Data-hungry Tasks Fine-tuning: We use the data-hungry tasks’ data from ACCN-INS for fine-tuning, resulting in the model named *TongGu-7B-trans*. **Data-efficient Tasks Fine-tuning:**

Algorithm 1 Redundancy-Aware Tuning (RAT)

Require: Model \mathcal{M} , Calibration Data \mathcal{D}_{old} , Training Data \mathcal{D}_{new} , Groups N

- 1: **for** $s \in \mathcal{D}_{old}$ **do**
 - 2: **for** $i = 1$ to $L - 1$ **do**
 - 3: $H_{i,t}, H_{i+1,t} \leftarrow \mathcal{M}(s)$
 - 4: $\text{CoS}_i \leftarrow \frac{1}{L} \sum_{t=1}^L \frac{H_{i,t} \cdot H_{i+1,t}}{\|H_{i,t}\|_2 \|H_{i+1,t}\|_2}$
 - 5: **end for**
 - 6: **end for**
 - 7: $\text{redundancy} \leftarrow \overline{\text{CoS}_i}$
 - 8: Divide layers into N groups: $\{G_1, \dots, G_N\}$
 - 9: **for** $g = 1$ to N **do**
 - 10: $l_g \leftarrow \arg \max_{l \in G_g} (\text{redundancy})$
 - 11: Fine-tune l_g on \mathcal{D}_{new} , keep others frozen
 - 12: **end for**
-

We proceed to use the data-efficient tasks’ data from ACCN-INS to fine-tune the *TongGu-7b-trans* model, cultivating the model’s capabilities into a broader range of CCU tasks. Moreover, we filter 10,000 dialogue samples between humans and AI assistants from ShareGPT (shareAI, 2023) as supplementary data, further enhancing the model’s conversational abilities. As a result, we obtained the final model named *TongGu-7B-Instruct*. To address the issue of catastrophic forgetting during the training process, we employ the proposed RAT method with N set to 8 throughout both stages. In the first stage, a subset of the incremental pre-training data serves as the calibration set. In the second stage, a subset of classical-to-modern Chinese translation corpus is used for the same purpose. More details can be found in Appendix B.

4.3 CCU-RAG

In knowledge-intensive CCU tasks, general-purpose LLMs and initial efforts in this field usually suffer from severe hallucinations. Recently, Retrieval-Augmented Generation (RAG) has been proven to be an effective solution in mitigating these hallucinations in LLMs (Gao et al., 2023; Zhao et al., 2024). Hence, we propose the CCU-RAG, a task-specific efficient RAG framework to enhance the veracity and reliability of the generated outputs from TongGu.

Firstly, knowledge-intensive samples were extracted from the instruction data, encompassing **source retrieval**, **author retrieval**, **previous sentence recitation**, **next sentence recitation**, **entire poem recitation**. Subsequently, these samples are reformatted into two types of data to simulate the

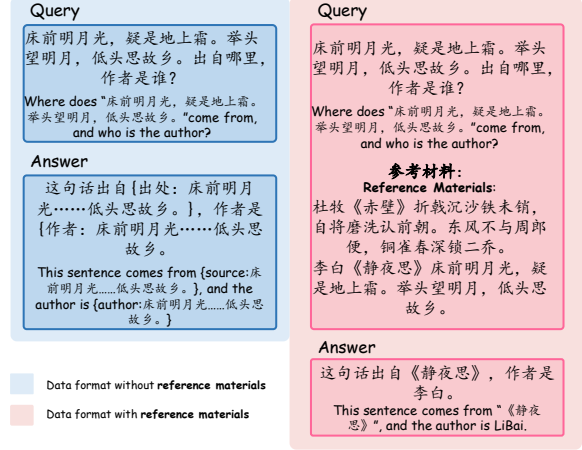


Figure 4: Two examples of reformatted knowledge-intensive tasks, with the difference being whether reference materials are provided.

two steps in RAG, as exemplified in Figure 4. One format retains the original query, with the response being reformulated as multi-level key-value pairs that support searching and retrieving. The other format involves appending *reference materials* to the original query, while maintaining the original response. The workflow of TongGu with the CCU-RAG system is illustrated in Figure 5. When TongGu receives a user query, if it is a knowledge-intensive task and lacks sufficient relevant knowledge, it generates a multi-level key-value pair to call the retrieval module. Then, the retrieved content is concatenated into the second instruction format and re-entered into the TongGu, enabling it to output more accurate answers. This judgment process is completed by the TongGu itself. Ultimately, these reformatted samples were utilized to replace the original samples in ACCN-INS, resulting in an enhanced retrieval-augmented instruction fine-tuning database.

It is worth noting that generating complete key-value pairs for lengthy sentences can be time-consuming. Therefore, we fine-tuned the model to focus solely on generating the beginning and ending fragments of the key-value pairs, using ellipses to replace excessively long intermediate text segments. The complete text in the context is used for retrieval based on the uncompleted text segment generated by the model. This approach is simple yet effective, significantly reducing the time required from user input to model response.

Table 3: Evaluation results for various LLMs based on performance on C³bench. * represents our reproduction. **Bold** indicates the best score, underline indicates the second best result.

Model	Classifications \uparrow	Retrieval \uparrow	NER \uparrow	Punctuation \uparrow	Translation \uparrow	Avg. \uparrow
Bloom-7B-Chunhua	39.62	13.36	34.70	62.19	11.27	33.23
Baichuan2-7B-Chat	37.00	18.36	63.25	53.96	13.70	37.15
Baichuan2-13B-Chat	44.26	17.79	46.67	65.11	12.45	37.26
ChatGLM2-6B	50.28	9.03	28.56	28.48	6.76	24.62
Qwen-7B-Chat	49.65	13.92	28.33	69.61	15.61	35.42
Qwen-14B-Chat	44.93	<u>25.90</u>	<u>66.72</u>	71.83	15.38	44.95
LLaMA2-Chinese-7B-Chat	18.78	3.20	12.62	34.73	4.24	14.71
LLaMA2-Chinese-13B-Chat	28.75	2.27	9.31	47.27	5.91	18.70
Moss-moon-003-SFT	15.07	15.84	28.90	58.39	13.35	26.30
GPT-3.5-turbo	50.65	7.36	63.83	61.34	11.94	39.02
GPT-4	53.88	13.71	63.87	67.31	12.09	42.17
ERNIE-bot-turbo	50.70	21.22	9.61	65.29	10.66	31.50
Spark-v3	51.61	21.83	53.81	<u>85.38</u>	<u>34.58</u>	<u>49.44</u>
abab5-chat	52.20	15.53	34.64	65.42	10.56	35.67
ChatGLM_Turbo	56.15	20.49	30.04	69.72	10.91	37.46
Xunzi-Qwen-Chat*	42.48	11.58	54.64	78.62	16.64	40.79
GLM4-9B-Chat*	52.99	19.73	48.15	70.96	17.26	41.82
Qwen2-7B-Instruct*	<u>56.22</u>	20.60	64.36	73.19	15.17	45.91
TongGu-7B-Instruct (Ours)	72.47	77.30	73.46	89.97	54.43	74.53

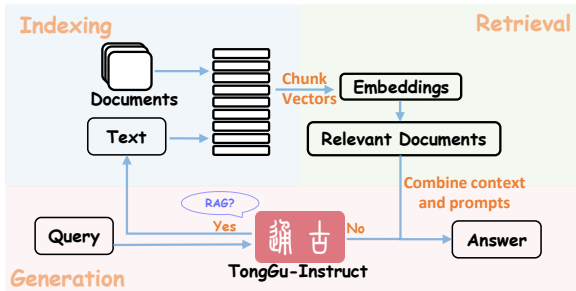


Figure 5: Workflow of TongGu response with CCU-RAG.

5 Experiments

In our experiments, we focus on evaluating TongGu’s capabilities in Classical Chinese Understanding (CCU) across three dimensions: understanding, generation, and knowledge.

5.1 Performance on C³bench

To evaluate the performance of TongGu on common classical Chinese tasks, we utilize the C³bench (Cao et al., 2024), a comprehensive classical Chinese benchmark designed for LLMs, covering ten domains and five common classical Chinese tasks.

The details of C³bench is presented in Appendix C, Table 10. We rigorously adhere to the settings outlined in the C³bench paper and conduct a zero-shot evaluation. The quantitative results of TongGu’s performance on the C³bench are pre-

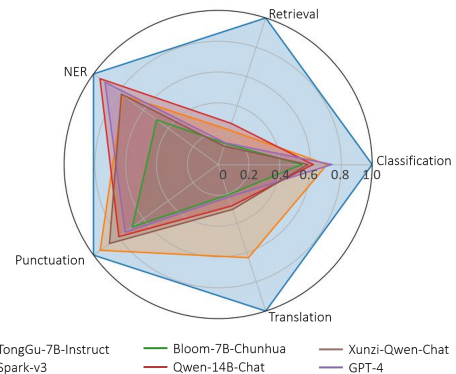


Figure 6: Radar charts of performances on C³bench. The values have been normalized to a 0-1 scale using the metrics of TongGu-7B-Instruct.

sented in Table 3, alongside the more radar charts in Figure 6. As confirmed by the outcomes, we can observe that TongGu outperforms existing LLMs across all five tasks, especially in knowledge-intensive retrieval task and data-hungry translation task.

5.2 Performance on Broader CCU Tasks

The ACCN-INS dataset covers a broader range of tasks than existing benchmarks, including tasks such as poetry creation and Flying flower order (a form of Chinese literary game requiring a poem containing a certain keyword). For a comprehensive evaluation, we generate a test set of 1,600 samples using the same process as the training set,

utilizing it only during model evaluation. The details of our test dataset are presented in Appendix C, Table 11. In our evaluation, we adopted a zero-shot approach, separately assessing knowledge-intensive tasks and Non-knowledge-intensive tasks, and conduct simultaneous testing of Baichuan2-7B-Chat (Baichuan, 2023) as a baseline.

Table 4: Comparison of Baichuan2 and TongGu on a wider range of classical Chinese tasks.

Task	Baichuan2-7B-Chat	TongGu-7B-Instruct
<i>Knowledge-intensive Tasks</i>		
	ACC ↑	ACC ↑
Source Retrieval	0.00	96.67
Author Retrieval	3.33	100.00
Previous Sentence Recitation	0.00	46.67
the Next Sentence Recitation	10.00	83.33
Entire Poem Recitation	0.00	96.67
<i>Non-knowledge-intensive Tasks</i>		
	PPL ↓	PPL ↓
Grammar	5.63	4.84
Ancient Poetry Writing	10.44	8.25
Couplet	26.63	43.25
Classical Chinese to Modern Chinese	60.00	10.75
Modern Chinese to Classical Chinese	21.75	20.75
Classical Chinese Writing	20.75	14.00
Poet Introduction	17.75	11.06
Analysis of Imagery	17.13	3.40
Knowledge of Sinology Q&A	9.06	8.75
Idiom	8.13	6.84
Riddle	13.38	12.75
Xiehoyu	10.75	27.50
Flying Flower Order	3.48	3.83
Named Entity Recognition	13.56	8.13
Punctuation	5.09	3.63
Topic Classification	9.81	7.16
Word Explanation	18.25	14.44
Reading Comprehension	2.22	1.88
Poetry Appreciation	12.56	10.75

For knowledge-intensive tasks, we utilize accuracy as the metric. For non-knowledge-intensive tasks, we use perplexity (PPL) as the metric, where the question and answer are concatenated and input into the model to compute the PPL. Table 4 presents the test results. TongGu-7B-Instruct outperforms Baichuan-7B-Chat in 21 out of 24 tasks, which substantiates the effectiveness of our incremental pre-training and the two-stage fine-tuning approach (RAT). More results can be found in Appendix D.

5.3 Ablation Study

Fine-tuning methods. We compared our proposed Redundancy-Aware Tuning (RAT) with two other fine-tuning methods: full-parameter fine-tuning (FT) and Low-Rank Adaptation (LoRA), using the C³bench. The results are summarized in Table 5. The results show that the vanilla FT method performs well in learning new tasks but suffers from catastrophic forgetting. The LoRA method mitigates catastrophic forgetting to some extent but struggles to adapt effectively to new tasks. In contrast, our proposed RAT method outperforms both FT and LoRA in terms of mitigating catastrophic forgetting and learning new tasks effectively. Ab-

lation results on a wider range of tasks can be found in Appendix E, Table 12.

RAT specifications. We investigate the impact of different values of N on the model performance used in the RAT method, as shown in Table 6. Setting N to 8 provides the best performance, hence we adopt it as the default strategy.

The effectiveness of CCU-RAG. We evaluated the impact of our proposed CCU-RAG method by comparing TongGu with and without CCU-RAG on the C³bench dataset, with results summarized in Table 4. The results reveal that the CCU-RAG method significantly improves TongGu’s performance on the knowledge-intensive task of source retrieval, without diminishing the performance of Non-knowledge-intensive tasks such as punctuation and named entity recognition. Results on a wider range of tasks are given in Appendix E, Table 13.

6 Conclusion

In this paper, we introduce TongGu, a new state-of-the-art LLM specifically for Classical Chinese Understanding (CCU). Our contributions include the development of the ACCN-INS dataset, which serves as the first publicly accessible CCU instruction dataset, and the introduction of innovative techniques such as Redundancy-Aware Tuning (RAT) and CCU-RAG (Retrieval-Augmented Generation). Through extensive experiments and evaluations, we have demonstrated TongGu’s superior performance in diverse CCU tasks, surpassing existing LLMs by a large margin in both knowledge-intensive and non-knowledge-intensive tasks. We believe that TongGu and the AC-INS dataset will serve as valuable resources for future endeavors in the CCU research community.

7 Limitations

TongGu’s performance heavily relies on the quality and quantity of the instruction-tuning dataset (ACCN-INS). The ACCN-INS dataset, while comprehensive, may not capture all variations in Classical Chinese texts. The RAT fine-tuning techniques, although effective, may still face challenges in mitigating catastrophic forgetting. Despite using CCU-RAG, the model may still produce hallucinations. Overcoming these limitations and expanding research in CCU will drive progress in understanding classical languages and cultural heritage.

Table 5: Ablation study of various training strategies.

Method	Data-hungry Tasks	Data-efficient Tasks					Avg. ↑
	Translation ↑	Classifications ↑	NER ↑	Punctuation ↑	Retrieval ↑	Avg. ↑	
FT	38.57	72.42	79.66	90.92	65.71	77.18	69.46
LoRA	49.13	42.43	18.29	89.31	33.04	45.77	46.44
RAT (Ours)	54.43	72.47	73.46	89.97	77.30	78.30	73.53

Table 6: Ablation study on the impact of different N used in RAT.

N	Data-hungry Tasks	Data-efficient Tasks					Avg. ↑
	Translation ↑	Classifications ↑	NER ↑	Punctuation ↑	Retrieval ↑	Avg. ↑	
4	54.47	66.24	69.02	89.91	79.78	76.24	71.88
8	54.43	72.47	73.46	89.97	77.30	78.30	73.53
16	50.48	74.00	74.83	87.93	54.75	72.88	68.40

Table 7: Ablation study on the effect of CCU-RAG.

Method	Non-knowledge-intensive Tasks					Knowledge-intensive Tasks	Avg. ↑
	Translation ↑	Classifications ↑	NER ↑	Punctuation ↑	Avg.	Retrieval ↑	
Ours	54.43	72.47	73.46	89.97	72.58	77.30	73.53
w/o RAG	53.16	72.43	73.89	88.53	72.00	21.03	61.81

8 Ethical and Social Implications

TongGu has been trained to focus on processing ancient texts related to Chinese culture. Despite a series of cleaning and review processes applied to the training data, there may still be factual errors present, which could lead TongGu, like other large language models, to generate misleading information or harmful content that contains factual errors. In the future, we will continue to fine-tune and release updated versions as we progress in addressing these issues.

Acknowledgments

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. 2024. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). In *Proc. ICLR*.

Baichuan. 2023. [Baichuan 2: Open large-scale language models](#). *arXiv preprint arXiv:2309.10305*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proc. ICML*, pages 2206–2240. PMLR.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proc. NeurIPS*, pages 1877–1901.

Jiahuan Cao, Yongxin Shi, Dezhi Peng, Yang Liu, and Lianwen Jin. 2024. [C³bench: A comprehensive classical Chinese understanding benchmark for large language models](#). *Preprint, arXiv:2405.17732*.

Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, and Vera Demberg. 2021. Time-Aware ancient Chinese text translation and inference. In *Proc. LChange*, pages 1–6.

Liu Chang, Wang Dongbo, Zhao Zhixiao, Hu Die, Wu Mengcheng, Lin Litao, Shen Si, Li Bin, Liu Jiangfeng, Zhang Hai, et al. 2023. SikuGPT: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *arXiv preprint arXiv:2304.07778*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PALM: Scaling language modeling with pathways. *JMLR*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *JMLR*, 25(70):1–53.

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*.

- Garychowcmu. 2019. Daizhige20. <https://github.com/garychowcmu/daizhige20>.
- Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. 2024. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proc. ICML*, pages 3929–3938. PMLR.
- Xiaowei Han, Lizhen Xu, and Feng Qiao. 2018. CNN-BiLSTM-CRF model for term extraction in Chinese corpus. In *Proc. WISA*, pages 267–274. Springer.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Proc. JMLR*, 24(251):1–43.
- Zongyuan Jiang, Jiapeng Wang, Jiahuan Cao, Xue Gao, and Lianwen Jin. 2023. Towards better translations from classical to modern Chinese: A new dataset and a new method. In *Proc. NLPCC*, pages 387–399. Springer.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proc. NeurIPS*, 33:9459–9474.
- Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.
- MOP-LIWU Community and MNBVC Team. 2023. Mnbvc: Massive never-ending bt vast chinese corpus. <https://github.com/esbatmop/MNBVC>.
- OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code LLaMA: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- shareAI. 2023. ShareGPT-Chinese-English-90k bilingual human-machine qa dataset. <https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k>.
- Together Computer. 2023. [Redpajama: An open source recipe to reproduce llama training dataset](https://github.com/togethercomputer/redpajama).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023b. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, et al. 2023. GujiBERT and GujiGPT: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.
- Wptoux. 2023. Bloom-7B-chunhua. <https://huggingface.co/wptoux/bloom-7b-chunhua>.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Further fine-tuning LLaMA on medical papers. *arXiv preprint arXiv:2304.14454*.
- Xunzi-LLM-of-Chinese-classics. 2024. XunziALLM. <https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM>.
- Peng Yu and Xin Wang. 2020. BERT-based named entity recognition in Chinese Twenty-Four Histories. In *Proc. WISA*, pages 289–301. Springer.
- Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on LLaMA model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for AI-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

A Dataset Details

In this section, we present the details of our dataset. Figure 7 illustrates an example of each task in the dataset. The dimension and quantity of each task are shown in Table 8. It is notable that the ACCN-INS dataset will be released under the CC-BY-NC-SA-4.0 license, and the ShareGPT (shareAI, 2023) dataset using in fine-tuning is under the apache-2.0 license.

Table 8: Details of each task in ACCN-INS.

Task	Source	Quantity
<i>comprehension Dimension</i>		
Classical Chinese to Modern Chinese	Labeled data	4,001,500
Modern Chinese to Classical Chinese	Labeled data	1,900
Named Entity Recognition	Labeled data	1,355
Punctuation	Labeled data	2,000
Topic Classification	Labeled data	1,200
Word Explanation	Unlabeled data	997
Reading Comprehension	Unlabeled data	363
Appreciation	Unlabeled data	600
Grammar	Unlabeled data	82
<i>Generation Dimension</i>		
Classical Chinese Writing	Unlabeled data	165
Ancient Poetry Writing	Unlabeled data	165
Couplet	Labeled data	800
<i>Knowledge Dimension</i>		
Knowledge of Sinology Q&A	Unlabeled data	1,000
Xiehouyu	Unlabeled data	1,000
Analysis of Imagery	Labeled data	1,000
Poet Introduction	Unlabeled data	946
Riddle	Unlabeled data	463
Idiom	Labeled data	1,000
Author Retrieval	Labeled data	600
Source Retrieval	Labeled data	600
Entire Poem Recitation	Labeled data	600
Previous Sentence Recitation	Labeled data	600
Next Sentence Recitation	Labeled data	600
Flying Flower Order	Labeled data	56
total		4,020,136

B Training Details

In this section, we present the training details of TongGu. The hyper-parameter settings of incremental pre-training and instruction tuning are shown in 9. All experiments are completed on 6 Nvidia A6000 GPUs.

C Evaluation Details

In this section, we detail the evaluation of TongGu on the C³bench and our test dataset. The results, including the number of tasks and the evaluation metrics, are summarized in Tables 10 and 11.

Table 9: Hyper-parameter settings in incremental pre-training and instruction-tuning.

Hyper parameter	Value	
	Incremental Pretraining	Instruction-tuning
Precision	bf16	bf16
Epoch	1	3
Batch size	576	192
Learning rate	2e-5	2e-6
Weight decay	0	0
Warmup ratio	0	0
LR scheduler type	cosine	cosine
Optimizer	AdamW	AdamW
β_1	0.9	0.9
β_2	0.999	0.999
Max length	2048	2048

Table 10: Details of C³bench.

Task	Metric	Quantity
Classification	Accuracy	10,000
Retrieval	Accuracy	10,000
NER	F1-score	10,000
Punctuation	F1-score	10,000
Translation	BLEU	10,000
Total	-	50,000

Table 11: Details of our test dataset.

Task	Metric	Quantity
Source Retrieval	Accuracy	30
Author Retrieval	Accuracy	30
Previous Sentence Recitation	Accuracy	30
Next Sentence Recitation	Accuracy	30
Entire Poem Recitation	Accuracy	30
Grammar	PPL	30
Ancient Poetry Writing	PPL	30
Couplet	PPL	100
Classical Chinese Writing	PPL	30
Classical Chinese to Modern Chinese	PPL	100
Modern Chinese to Classical Chinese	PPL	100
Poet Introduction	PPL	50
Analysis of Imagery	PPL	100
Knowledge of Sinology Q&A	PPL	100
Idiom	PPL	100
Riddle	PPL	50
Xiehouyu	PPL	100
Flying Flower Order	PPL	30
Named Entity Recognition	PPL	100
Punctuation	PPL	100
Topic Classification	PPL	100
Word Explanation	PPL	100
Reading Comprehension	PPL	30
Poetry Appreciation	PPL	100
Total	-	1,600

Table 12: Performance comparison of various training strategies on a wider range of classical Chinese tasks. Cells in yellow represent the retrieval tasks.

Task	FT	LoRA	RAT
<i>Data-hungry Tasks</i>			
Classical Chinese to Modern Chinese	12.75	<u>11.06</u>	10.75
<i>Data-Efficient Tasks</i>			
Source Retrieval	96.67	0.00	96.67
Author Retrieval	<u>53.33</u>	6.67	100.00
Previous Sentence Recitation	60.00	33.33	<u>46.67</u>
Next Sentence Recitation	83.33	36.67	83.33
Entire Poem Recitation	96.67	66.67	96.67
Grammar	<u>5.10</u>	5.25	4.84
Ancient Poetry Writing	7.34	12.18	<u>8.25</u>
Couplet	33.00	45.75	<u>43.25</u>
Modern Chinese to Classical Chinese	<u>18.88</u>	18.00	20.75
Classical Chinese Writing	13.75	18.25	<u>14.00</u>
Poet Introduction	8.94	13.56	<u>11.06</u>
Analysis of Imagery	4.44	<u>3.63</u>	3.41
Knowledge of Sinology Q&A	<u>9.31</u>	9.50	8.75
Idiom	6.53	7.09	<u>6.84</u>
Riddle	13.81	<u>13.38</u>	12.75
Xiehouyu	18.88	<u>24.25</u>	27.50
Flying Flower Order	3.30	4.22	<u>3.83</u>
Named Entity Recognition	<u>7.87</u>	7.28	8.13
Punctuation	3.65	3.63	3.63
Topic Classification	<u>7.34</u>	8.41	7.16
Word Explanation	15.63	13.38	<u>14.44</u>
Reading Comprehension	2.11	<u>1.89</u>	1.88
Poetry Appreciation	13.19	<u>12.93</u>	12.75

D Sample Responses from Different Models

For each dimension, we present a task example, including the user’s query, as well as the responses from Baihuan2-7B-Chat, GPT-3.5, and our TongGu model. It can be seen from Figure 8 to 10 that our TongGu model demonstrates excellent performance across the three dimensions of understanding, generation, and knowledge, proving the effectiveness of our data and training methods.

E Ablation Study Details

In this section, we present the ablation study in our test dataset. Table 12 shows the performance of our RAT method and other training strategies, from which we can see that our RAT method not only alleviates catastrophic forgetting in data-hungry tasks but also generalizes well to data-efficient tasks. Table 13 shown the ablation study on the effect of CCU-RAG, from which we can summarize that CCU-RAR significantly enhances the performance of knowledge-intensive tasks without compromising the Non-knowledge-intensive tasks.

Table 13: Ablation study on the effect of CCU-RAG on a wider range of classical Chinese tasks.

Task	Ours	w/o RAG
<i>Knowledge-intensive Tasks</i>		
	ACC ↑	ACC ↑
Source Retrieval	96.67	3.33
Author Retrieval	100.00	30.00
Previous Sentence Recitation	46.67	0.00
Next Sentence Recitation	83.33	33.33
Entire Poem Recitation	96.67	0.00
<i>Non-knowledge-intensive Tasks</i>		
	PPL ↓	PPL ↓
Grammar	4.84	4.89
Ancient Poetry Writing	8.25	8.25
Couplet	43.25	42.50
Classical Chinese to Modern Chinese	10.75	10.75
Modern Chinese to Classical Chinese	20.75	21.75
Classical Chinese Writing	14.00	14.38
Poet Introduction	11.06	10.44
Analysis of Imagery	3.41	3.44
Knowledge of Sinology Q&A	8.75	8.75
Idiom	6.84	6.98
Riddle	12.75	12.75
Xiehouyu	27.50	25.75
Flying Flower Order	3.83	3.85
Named Entity Recognition	8.13	7.88
Punctuation	3.63	3.52
Topic Classification	7.16	6.78
Word Explanation	14.44	14.00
Reading Comprehension	1.88	1.52
Poetry Appreciation	12.75	12.75

<p>文言文-白话文翻译 Classical Chinese to Modern Chinese</p>	<p>翻译为白话文(Translate into modern Chinese): 罢陕西五路西蜀四川行尚书省,以也送峇儿行四川尚书省掌于兴元,京兆等路直隶尚书省。改宋军于渭河。</p>	<p>撤销陕西五路、西蜀、四川的行尚书省,派也送峇儿在兴元执掌四川行尚书省的事务,把京兆等路直接隶属于尚书省。在渭河击败宋军。</p>
<p>白话文-文言文翻译 Modern Chinese to Classical Chinese</p>	<p>翻译成文言文(Translate into classical Chinese): 有一些人是很容易模仿的,但是王杰的一切个性、特色与成绩,却从没有人可以取代。</p>	<p>有若干人,极易模仿,而王杰之一切性质、特征与成就,竟无人得以代之。</p>
<p>命名实体识别 Named Entity Recognition</p>	<p>请从以下文言文中提取出实体(Please extract the entities from the following classical Chinese text): 戊寅,募兵于通州、天津、宣府、大同。</p>	<p>该句子中的命名实体有(The named entities in the sentence are): 戊寅、通州、天津、宣府、大同。</p>
<p>句读 Punctuation</p>	<p>给下列句子添加标点符号(Adding punctuation to the following sentence): 折戟沉沙铁未销自将磨洗认前朝东风不与周郎便铜雀春深锁二乔</p>	<p>折戟沉沙铁未销,自将磨洗认前朝。东风不与周郎便,铜雀春深锁二乔。</p>
<p>主题分类 Topic Classification</p>	<p>从道、法、诗、农、佛中选择一项作为下面句子的类(Choose one category from Taoism, Law, Poetry, Agriculture, and Buddhism for the following sentence): 空劳苦,唐失身命,未曾善心为于法也。</p>	<p>句子的类别为: 佛(The category of the sentence is: Buddhism)</p>
<p>字词解释 Word Explanation</p>	<p>请问“九成吹玉珥,百尺上瑶台”中的“珥”是什么意思?(Could you please explain the meaning of '珥' in the phrase '九成吹玉珥,百尺上瑶台'?)</p>	<p>这里的“珥”,是古代的一种竖吹乐器,类似现代的竖吹笛子。(The term '珥' here refers to an ancient vertical wind instrument, similar to a modern flute.)</p>
<p>阅读理解 Reading Comprehension</p>	<p>谢太傅寒雪日内集,与儿女讲论文义。俄而雪骤,公欣然曰:“白雪纷纷何所似?”兄子胡儿曰:“撒盐空中差可拟。”兄女曰:“未若柳絮因风起。”公大笑乐。即公大兄无奕女,左将军王凝之妻也。整篇文言文《咏雪》主要描述了什么场景?(What scene does the entire 《咏雪》 mainly depict?)</p>	<p>这篇文言文主要描述了谢太傅在一个下雪天与儿女们聚会,一起讨论诗文。当雪下得越来越大时,他提出了一个关于雪花如何比喻的问题,引发了儿女们的讨论。整篇文章以咏雪为题材,展示了家庭成员之间的文化交流和家庭和谐氛围。(This classical Chinese text primarily describes...)</p>
<p>诗词赏析 Poetry Appreciation</p>	<p>《夜归鹿门山歌》是如何描绘景色的?(How does the poem 《夜归鹿门山歌》 depict the scenery?)</p>	<p>整首诗按照时空顺序,分别写了江边和山中两个场景,先动后静,以动衬静,写出鹿门清幽的景色。(This poem...)</p>
<p>文言文写作 Classical Chinese Writing</p>	<p>以图书馆学习为主题,用文言文写一篇记叙文。(Write a narrative essay in classical Chinese with the theme of studying in a library.)</p>	<p>钟声未绝,余已攘袖而起。匆匆步入校门,走过回廊,直抵图书馆。推门而入,万籁俱静,惟余笔耕墨耘之响。借一束阳光,温煦砚池;一番耕耘,透写纸上。有得经典解惑,有感诸子为人生指南;亦有蹉跎岁月,奋发图强。我已淹留数载,虽未尽窥全豹,已识其门径。况复书阁无涯,渺然迷失于汪洋恣肆,亦自快哉。倚荏苒书海,有朋来扶持共游,或将陶醉其中,绝不复返。盖读万卷高复杯水,努力方显英雄本色。纵横驰骋,傲然凌人,从此不夫英雄偶风。</p>
<p>诗词创作 Ancient Poetry Writing</p>	<p>请以“青春”为主题,《如梦令》为词牌名创作一首词。(Please create a poem with the theme of 'Youth' and using 《如梦令》 as the poetic form.)</p>	<p>《如梦令·青春》年少轻狂如梦,壮志凌云谁共?岁月逝无痕,唯有情怀依旧。回首,回首,往事如烟飘散。</p>
<p>对联 Couplet</p>	<p>对对联(Matching couplets): 八音金石丝竹鲍土草木</p>	<p>九宫体生衍杜景天惊开</p>
<p>诗词默写 Poetry Dictation</p>	<p>落木千山天远大, _____。</p>	<p>澄江一道月分明</p>
<p>歇后语 Xiehouyu</p>	<p>请补充下面歇后语(Please complete the following two-part allegorical saying): 西天取经</p>	<p>任重道远</p>
<p>意象解释 Analysis of Imagery</p>	<p>在古典文学里,“晚来”这个词一般用来表达什么意思?(In classical literature, what does the term “晚来” generally express?)</p>	<p>傍晚;入夜。唐杜甫《题郑县亭子》诗:“更欲题诗满青竹,晚来幽独恐伤神。”</p>
<p>诗人介绍 Poet Introduction</p>	<p>苏轼有哪些代表作?(What are some of Su Shi's representative works?)</p>	<p>苏轼的代表作有(Su Shi's representative works include)《水调歌头·明月几时有》、《念奴娇·赤壁怀古》、《江城子·密州出猎》等。</p>
<p>成语解释 Idiom</p>	<p>“文笔流畅”是什么意思?(What does “文笔流畅” mean?)</p>	<p>“文笔流畅”意味着(“文笔流畅” means)文字通顺,读起来不费力。</p>
<p>谜语 Riddle</p>	<p>猜谜语:七十二小时,请问这个谜语的谜底是哪一个汉字?(Riddle: Seventy-two hours, what is the answer to this riddle in terms of a Chinese character?)</p>	<p>晶。解释:七十二小时是三天,三天即为三日,‘晶’字可看作三个‘日’字组合而成。(“晶” Explanation: Seventy-two hours is three days, and “three days” can be interpreted as “three suns”. The character ‘晶’ is composed of three ‘sun’ (日) characters.)</p>
<p>国学常识 Knowledge of Sinology Q&A</p>	<p>18世纪,七巧板传到国外,西方人把七巧板称作什么?(In the 18th century, the Tangram was introduced to foreign countries. What did Westerners call the Tangram?)</p>	<p>唐图(Tang picture).</p>
<p>创作背景 Background Introduction</p>	<p>《登金陵凤凰台》是在什么情况下创作的?(Under what circumstances was 《登金陵凤凰台》 created?)</p>	<p>《登金陵凤凰台》的创作背景说法不一,一说是李白被排挤离开长安,南游金陵时所作;一说是作者流放夜郎遇赦返回后所作;也有人称是李白游览黄鹤楼,想与崔颢的《黄鹤楼》争胜。(There are various accounts regarding the background of the creation of 《登金陵凤凰台》. One theory suggests that it was written when Li Bai was ostracized and left Chang'an, traveling to Jinling; another theory posits that it was composed after the author was pardoned and returned from exile in Ye Lang; some also claim it was written by Li Bai while visiting the Yellow Crane Tower, with the intention of rivaling Cui Hao's 《黄鹤楼》.)</p>
<p>飞花令 Flying Flower Order</p>	<p>请背一首含有“春”字的古诗词。(Please recite a classical Chinese poem or verse that contains the character “春”.)</p>	<p>春眠不觉晓,处处闻啼鸟。——孟浩然《春晓》</p>
<p>出处检索 Source Retrieval</p>	<p>“天生我材必有用,千金散尽还复来”出自哪里?(Where does “天生我材必有用,千金散尽还复来” come from?)</p>	<p>这句话出自(It comes from)《将进酒》。</p>
<p>作者检索 Author Retrieval</p>	<p>“天生我材必有用,千金散尽还复来”的作者是谁?(Who is the author of “天生我材必有用,千金散尽还复来”?)</p>	<p>“天生我材必有用,千金散尽还复来”的作者是李白。(The author of “天生我材必有用,千金散尽还复来” is Li Bai.)</p>

Figure 7: Example of each task in ACCN-INS.





	<p>Question: 创作一首藏头诗，每句的第一个字合起来是“星辰大海”。 Create an acrostic poem where the first character of each line spells out “星辰大海”。</p>
	<p>GPT-3-5: 星辰璀璨世界启，星光闪耀夜空辉。星际之门虚掩启，星云缭绕宇宙飞。 辰光照耀万物生，辰宇无垠充满晖。大气漫游海外浩，大宇宙中蕴奥秘。 海洋浩渺涵宇宙，海市蜃楼幻影姿。海浪汹涌波澜起，海底珊瑚世界齐。✘</p>
	<p>Baichuan2-7B-Chat: 星辰闪烁映银河，海天一色接天涯。大海浩渺藏奥秘，宇宙无垠探无穷。✘</p>
	<p>Tonggu-7B-sft: 星辉闪烁映苍穹，辰宿列张耀长空。大地无垠藏奥秘，海天一色任西东。✔</p>

Figure 8: Examples of responses to generation dimension tasks from different models.





	<p>Question: 请指出句子“谨拜表以闻”中的省略成分，并补全句子。 Please identify the omitted elements in the sentence “谨拜表以闻” and complete the sentence.</p>
	<p>GPT-3-5: 句子“谨拜表以闻”是一种古代的书信结尾用语，常见于中国传统文化中。在该句中，存在省略成分。推测完整的句子可能是：“谨拜表以闻，敬祈阁下安康”。这句话的完整含义是：“谨以拜表的方式向您报告，恭敬地祈祷您身体健康。”在这个句子中，“敬祈阁下安康”是对收信人的尊敬和祝福，往往用于书信结尾。 The phrase “谨拜表以闻” is a form of closing used in ancient letters, commonly found in traditional Chinese culture. There are omitted elements in this sentence. It is speculated that the complete sentence might be: “谨拜表以闻，敬祈阁下安康”. The full meaning of this sentence is: “I respectfully report to you through this letter, and sincerely wish you good health.” In this sentence, “敬祈阁下安康” is a sign of respect and blessing to the recipient, often used at the end of letters. ✘</p>
	<p>Baichuan2-7B-Chat: 在这个句子中，省略的成分是“臣”，补全后的句子应该是：“谨拜表以闻，臣”。 In this sentence, the omitted element is “臣”. The completed sentence should be: “谨拜表以闻，臣”. ✘</p>
	<p>Tonggu-7B-sft: 该句省略了宾语，应补全为“谨拜表以（之）闻”。意为“恭敬地呈上此表来使陛下知道这件事”。 The sentence omits the object and should be completed as “谨拜表以（之）闻”. It means “I respectfully present this petition to make Your Majesty aware of this matter.” ✔</p>

Figure 9: Examples of responses to understanding dimension tasks from different models.

 **Question:**
在我国风俗中，常常避讳73和84这两个岁数，因为这是两位历史人物去世的年龄，他们是谁？
In Chinese customs, the ages of 73 and 84 are often taboo because these are the ages at which two historical figures passed away. Who are they?

 **GPT-3.5:**
我很抱歉，但我无法回答你提供的关于历史人物去世年龄的具体问题。作为一个语言模型，我的知识截止日期是2021年，而我没有关于中国历史上特定人物去世年龄的详细信息。在中国文化中，有时会出于一些迷信或风俗习惯而避讳某些数字或岁数，但这种习俗可能因地区和个人信仰而有所不同。如果你对中国历史和文化的特定方面有更具体的问题，我将尽力提供帮助。
I apologize, but I cannot answer the specific question you provided about the age at which historical figures passed away. As a language model, my knowledge is up to date until 2021, and I do not have detailed information on the specific ages at which particular figures in Chinese history died. In Chinese culture, there are times when certain numbers or ages are avoided due to superstitions or customs, but these practices may vary by region and personal beliefs. If you have more specific questions about certain aspects of Chinese history and culture, I will do my best to assist you. ❌

 **Baichuan2-7B-Chat:**
在中国传统文化中，对某些数字的避讳源于古代对自然现象的理解和对神灵的敬畏。例如，“七”在佛教中被认为是地狱的数目，因此人们通常避免使用“七”这个数字。而“八”则因为与“发”谐音，被认为是不吉利的，所以人们也避免使用。然而，对于“七十三”和“八十四”这两个年龄的避讳，并没有明确的宗教或文化原因。这可能是由于这两个数字在数学上的特殊性质，使得它们在某些情况下容易被强调或误解。此外，人们可能还受到一些民间传说和故事的影响，这些故事中的人物通常在73岁或84岁时遭遇不幸。总之，对“七十三”和“八十四”这两个年龄的避讳，可能是由于多种原因的综合作用，但并没有确凿的宗教或文化依据。
In traditional Chinese culture, the taboo against certain numbers stems from ancient understandings of natural phenomena and reverence for deities. For instance, the number "seven" is considered in Buddhism to be the number of hell, so people generally avoid using the number "seven". The number "eight" is also avoided because it is a homophone for "lose", which is considered inauspicious. However, there is no clear religious or cultural reason for the taboo against the ages of "seventy-three" and "eighty-four". This may be due to the special mathematical properties of these numbers, which can sometimes be emphasized or misunderstood in certain contexts. In addition, people may also be influenced by some folk tales and stories, in which characters often encounter misfortune at the ages of 73 or 84. In summary, the taboo against the ages of "seventy-three" and "eighty-four" may be due to a combination of various reasons, but there is no definitive religious or cultural basis. ❌

 **Tonggu-7B-sft:**
在我国风俗中，73和84这两个岁数之所以常被避讳，是因为它们分别对应着两位著名历史人物的去世年龄。这两位历史人物是孔子和孟子，他们对中国文化有着深远的影响，并在儒家思想中占有重要地位。
In our customs, the ages of 73 and 84 are often avoided because they correspond to the ages at which two famous historical figures passed away. These two historical figures are Confucius and Mencius, who have had a profound influence on Chinese culture and hold an important position in Confucian thought. ✅

Figure 10: Examples of responses to knowledge dimension tasks from different models.