Book2QA: A Framework for Integrating LLMs to Generate High-quality QA Data from Textbooks

Anonymous EMNLP submission

Abstract

The scarcity of question-answering data is one of the main bottlenecks restricting the development of intelligent education systems. In this paper, we proposes a new method called Book2QA, which integrates multiple medium-scale language models (e.g., 6B/13B) to cost-effectively generate high-quality question-answering data from textbook content. The Book2QA framework includes three main steps: book data preprocessing, question generation with subsequent filtering, and answer generation with subsequent filtering. Our experimental results demonstrate the fine-tuned model's performance in real scenarios, highlighting the effectiveness of the Book2QA method. Automatic evaluation and advanced LLM evaluation show that data generated by Book2QA can match or surpass data from models with hundreds of billions of parameters. We open-source our data and code at https://anonymous.4open.science/r/Book2QA-F795.

1 Introduction

Building educational question-answering robots is highly significant but also extremely challenging in the education field, typically requiring educational datasets for model pre-training or fine-tuning. Recent studies have shown that directly fine-tuning with unlabeled data leads to performance degradation (Li et al., 2021), but supervised fine-tuning (SFT) on high-quality datasets can achieve performance breakthroughs in downstream tasks (Ouyang et al., 2022). Due to privacy and security issues, obtaining high-quality educational datasets is very difficult (Ouyang et al., 2022), and the datasets and benchmarks used for educational applications vary greatly in scope and purpose (Wang et al., 2024b). Additionally, there are cost concerns (Kasneci et al., 2023), which further complicate the construction of educational chatbots. Currently, content-based generation is a promising solution. Studies have shown that using advanced LLMs to generate fine-tuning datasets from book content yields good results (Wang et al., 2024a). However, this approach is costly and lacks diversity in the generated data. Therefore,we propose a framework (Book2QA) that addresses the shortage of high-quality question-answer data in the education sector by cost-effectively integrating the capabilities of multiple medium-sized language models (6B/13B parameters). Based on this framework, we have generated a new dataset for fine-tuning question-answering robots using textbook data.

Our framework generates question-answering data from book content and uses a fine-tuned student model for data filtering with IFD score (Li et al., 2024c) and the reverse IFD (r-IFD) score (Li et al., 2024a). The framework includes three main steps: book data preprocessing, question generation with subsequent filtering, and answer generation with subsequent filtering(as shown in Figure 1). In the process of generating question-answer pairs, the formulation of questions is particularly crucial(Sultan et al., 2020). In the process of generating question-answer pairs, the questions should both cover the details of the book and appropriately extend the content. Synthetic datasets generated using simple prompts exhibit significant bias and lack diversity (Yu et al., 2023). Therefore, during the questioning phase, we designed a set of prompting methods based on Bloom's taxonomy of educational objectives (Anderson et al., 2000) and integrated various levels of information to ensure that the large language model can generate diverse and high-quality questions.

In the field of education, questions and answers should not only contain key information from textbooks but also include additional rich information. Thus, evaluating this task is very challenging. Therefore, we designed two evaluation methods: automatic evaluation and advanced large language

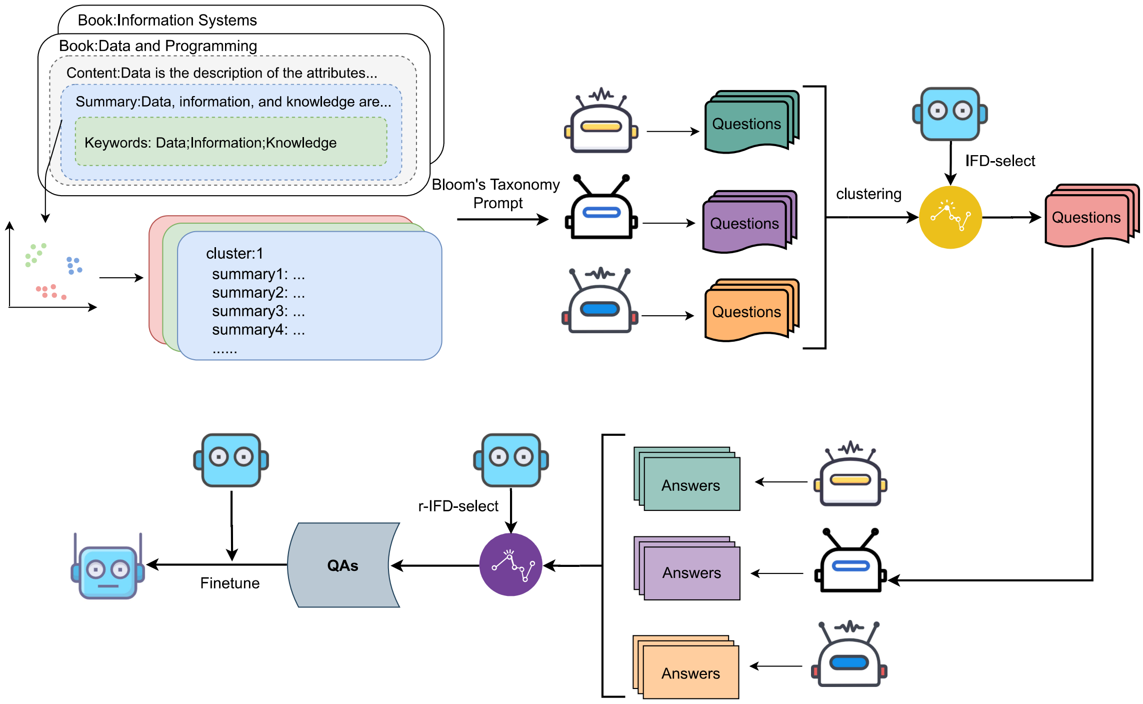

Figure 1: **Book2QA framework.** First, it extracts keywords from original paragraphs and creates summaries for semantic clustering. Second, multiple language models generate questions guided by Bloom's taxonomy, which are then clustered and filtered by IFD scores. Finally, answers are generated by three models, and the most understandable response is selected based on r-IFD scores.

model evaluation. The results show that the data generated using Book2QA scored highest in most metrics, as shown in Section 5.2. The models fine-tuned with the generated data exhibited good generalization performance in cross-domain evaluations. Sections 5.3 and 5.4 show that the Book2QA framework performs well in both question and answer generation. Additionally, the data generated by multiple medium-sized language models can rival or even surpass the data generated by large models with hundreds of billions of parameters (Section 5.5).Our contributions include:

- We propose the Book2QA framework, which integrates multiple medium-scale language models to provide a low-cost and efficient method for generating educational question-and-answer data.
- We developed a hierarchical prompting strategy based on Bloom's taxonomy to enhance the depth and breadth of the Q&A data.
- We fine-tuned the model on the generated data, and the results indicate that the fine-tuned model not only improves the quality of answer generation but also enhances its performance in practical applications.

2 Related Work

2.1 Data Synthesis

Recently, knowledge distillation (KD) has been proven effective in enhancing model capabilities (Xu et al., 2024). Large language models generate data to train smaller models, acting as a form of weak distillation that improves the performance of the smaller models (West et al., 2022). Introducing student models and using IFD and r-IFD metrics during distillation can enhance large language models' performance in instruction fine-tuning tasks(Li et al., 2024a). IFD scores from smaller models can filter training data for larger models. Different sized language models show high consistency in perceiving instruction difficulty(Li et al., 2024b). Additionally, study have focused on constructing data generators to train language models specifically designed for generating data in particular formats for knowledge distillation (Nayak et al., 2024).

Additionally, research shows that state-of-the-art models can feasibly generate high-quality content-grounded datasets (Yehudai et al., 2024), aligning with our goal of creating content-based question-answering pairs. While this method enhances question and answer quality, it is costly. Our approach combines content-based data generation from books with data filtering methods from knowledge distillation, rather than relying solely on distillation or stronger models for synthetic data generation.

2.2 Integration of Language Models

As more large language models emerge, research on integrating their outputs has gained attention. Studies show that dynamically integrating multiple LLMs produces outputs that better align with human feedback(Jiang et al., 2023). Integrating smaller language models is increasingly recognized. Combining three medium-sized models (6B/13B parameters) can rival large models like ChatGPT (175B+ parameters) and even surpass them in some performance metrics (Lu et al., 2024). Inspired by this, our approach proposes a method for data synthesis by integrating the outputs of multiple small to medium-sized language models.

3 Preliminaries

In our data generation framework, we use two key metrics: the IFD score and the r-IFD score. The IFD score shows how much adding instructions improves the chance of getting a response, with higher scores indicating harder instructions that the model needs to learn. The r-IFD score shows how well the response helps predict the instruction, with lower scores indicating that the model can easily infer the instruction, meaning the response and sample are suitable for the model.

3.1 Log-Likelihood

Log-likelihood is a way to measure the probability of a model generating a particular sequence. For a given sequence y and context x, its log-likelihood $L(y \mid x)$ can be expressed as:

$$L(y \mid x) = \sum_{i=1}^{n} \log P(y_i \mid y_{1:i-1}, x) \quad (1)$$

where $P(y_i \mid y_{1:i-1}, x)$ denotes the probability of generating the word y_i given the context x and the previously generated words $y_{1:i-1}$.

3.2 Instruction Following Difficulty

The IFD score measures how difficult it is for the student model to generate responses based on given instructions. It compares the perplexity of the model's response with and without the instruction. A higher IFD score means the instruction is more challenging and informative. Mathematically, IFD is defined as:

$$IFD(y|x) = \frac{ppl(y|x)}{ppl(y)} = e^{L(y|x)-L(y)} \quad (2)$$

Where $\text{ppl}(y|x)$ is the perplexity of the model generating a response given the instruction, and $\text{ppl}(y)$ is the perplexity of the response alone. $L(y|x)$ and $L(y)$ are the respective log-likelihoods, where $L(y)$ denotes the log-likelihood of generating the response y without any instruction.

3.3 Reverse Instruction Following Difficulty

The r-IFD score evaluates the potential to deduce the original instruction from the response. Specifically, r-IFD evaluates the feasibility of the response by comparing:

- The perplexity of the instruction given the response $\text{ppl}(x \mid y')$: This is the perplexity of the model generating the instruction x given the response y'.
- The perplexity of the instruction $\text{ppl}(x)$: This is the perplexity of the model generating the instruction x without any context.

Mathematically, r-IFD is defined as:

$$r-IFD(x \mid y) = \frac{ppl(x \mid y')}{ppl(x)} = e^{L(x|y')-L(x)} \quad (3)$$

$L(x \mid y')$ is the log-likelihood of generating the instruction x given the response y', and $L(x)$ is the log-likelihood of generating the instruction x without a response.

Integrating IFD and r-IFD allows for a comprehensive assessment of the instruction-response pair quality, ensuring that refined data better aligns with the learning capabilities and objectives of the student model. These two metrics jointly help in selecting and optimizing training data, enabling the model to perform more effectively and accurately in practical applications.

4 BOOK2QA: Generating High-Quality QA Data from Textbooks

Algorithm 1 shows our dataset generation method.

4.1 Textbook Data Preprocessing

We have developed a tool that converts PDF textbooks into structured JSON data J and segments the text into paragraphs, referred to as p. Each paragraph p within J is processed to generate a summary s_p and a set of n keywords K_p, enhancing the data's semantic richness.

$$s_p = \text{Summarize}(p)$$

$$K_p = \text{ExtractKeywords}(p, n)$$

We then cluster these summaries into a set C, providing enriched context for the subsequent QA pair generation:

$$C = \text{Cluster}(\{s_p | p \in J\})$$

Before processing, we initialize three datasets:

- D_{RETR} to store retrieved questions from clustering.
- D_{SYNTH} to hold the synthesized final QAs.
- D_{QGEN} to accumulate all generated questions.

Each dataset is initially empty, ensuring a clean slate for data processing.

4.2 Question Generation and Selection

Utilizing several medium-scale LLMs, we generate questions based on the contextual data structured through the taxonomy B (Bloom's Taxonomy). Each question generation step involves:

$$Q = \bigcup_{b \in B} \{m(\text{prompt}_b) | m \in M\}$$

where M represents the set of LLMs and B the cognitive levels from Bloom's Taxonomy. The prompts are dynamically generated based on three strategies, leveraging:

- Only the original text (o).
- A combination of the original text with a set of summaries from another clusters ($o + s$).
- Integration of the original text with keywords ($o + K$).

Algorithm 1 BOOK2QA

Require: Structured JSON textbook data (J), number of keywords per paragraph (n), number of clusters (k)
Ensure: High-quality QA pairs dataset (D)

Declaration: Let B denote Bloom's Taxonomy for cognitive levels.
Initialize: $D_{\text{RETR}} \leftarrow \emptyset$, $D_{\text{SYNTH}} \leftarrow \emptyset$, $D_{\text{QGEN}} \leftarrow \emptyset$
Textbook Data Preprocessing
for each paragraph p in J **do**
  $s_p \leftarrow \text{Summarize}(p)$
  $K_p \leftarrow \text{ExtractKeywords}(p, n)$
end for
$C \leftarrow \text{Cluster}(\{s_p\})$
Question Generation
$M \leftarrow$ {LLM1, LLM2, LLM3}
for each paragraph p in J **do**
  for each info in {"o", "o + s", "o + K"} **do**
    for each m in M **do**
      $prompt \leftarrow \text{Build}(p, info, B)$
      $Q_{pm} \leftarrow m.\text{QG}(prompt)$
      $D_{\text{QGEN}} \leftarrow D_{\text{QGEN}} \cup \{Q_{pm}\}$
    end for
  end for
end for
Perform k-means Clustering
$D_{\text{CLUST}} \leftarrow \text{k-means}(D_{\text{QGEN}}, k)$
for each cluster in D_{CLUST} **do**
  $Q^* \leftarrow \arg\max_{Q \in \text{cluster}}(\text{IFD}(Q))$
  $D_{\text{RETR}} \leftarrow D_{\text{RETR}} \cup Q^*$
end for
Answer Generation and Selection
for each Q^* in D_{RETR} **do**
  for each m in M **do**
    $A_{Q^*m} \leftarrow m.\text{AG}(Q^*, C)$
    $score_{Q^*m} \leftarrow \text{Calculate\_rIFD}(A_{Q^*m})$
  end for
  $A^*_{Q^*} \leftarrow \arg\min_{A_{Q^*m}}(score_{Q^*m})$
  $D_{\text{SYNTH}} \leftarrow D_{\text{SYNTH}} \cup (Q^*, A^*_{Q^*})$
end for
Combine and Output
return D_{SYNTH}

Post-question generation, we perform k-means clustering on D_{QGEN} to organize questions into k distinct clusters. From each cluster, we select the question with the highest IFD score, optimizing for question quality and relevance:

$$Q^* = \arg\max_{Q \in \text{cluster}}(\text{IFD}(Q))$$

4.3 Answer Generation and Selection

For each selected question Q^*, every model in M generates an answer, and we evaluate these answers based on their r-IFD scores to determine the most suitable:

$$A^*_{Q^*} = \arg\min_{A_{Q^*m}} \text{r-IFD}(A_{Q^*m})$$

where A_{Q^*m} denotes the answer produced by model m for Q^*. The lowest r-IFD score indicates the highest clarity and relevance.

Finally, the D_{SYNTH} dataset, which contains each question paired with its optimal answer, is returned.

These detailed steps from the algorithm ensure a systematic approach to generating a high-quality question-answering dataset from textbooks.

5 Experiment

In the experimental section, we used two textbooks, "Information Systems" and "Data and Programming" (see Appendix A), and utilized multiple medium-scale large language models, Baichuan13b-chat, Qwen7b-chat, and Internlm7b-chat to generate data for fine-tuning datasets(see Appendix Appendix E.1, our dataset termed book2qa_sft.) and conducted extensive experiments on these datasets.

5.1 Experimental Setup

Model We consider fine-tuning a large language model that has undergone pre-training and multi-turn dialogue alignment: Baichuan7B-Chat. This model, based on the Transformer architecture, is a 7-billion parameter model trained on approximately 1.2 trillion tokens, supporting both Chinese and English, with a context window length of 4096 (Yang et al., 2023).

Training Details We use LoRA (Hu et al., 2021) to fine-tune the language model on book and generated datasets. All models are trained for 1 epoch, with the lora_rank set to 8, lora_alpha set to 16, and the learning rate set to 5e-5.We used an A100 GPU to train the model, and the training time on the fine-tuning dataset mentioned in the paper is approximately 0.5 to 2 hours.

Evaluation Methods We will evaluate the quality of the generated data from two perspectives: automatic evaluation and advanced language model evaluation.

In automatic evaluation, we use BertScore (Zhang et al., 2020), ROUGE (Lin, 2004), IFD, r-IFD, text length, and Entropy to assess the generated dataset. BertScore and ROUGE are used to evaluate the semantic similarity between two texts, assessing the similarity between the generated text and the original paragraphs, as well as between questions and answers. IFD and r-IFD are used to evaluate the suitability of the QA data for student models. Text length is also an effective indicator of the quality of generated data, with studies showing that longer responses are highly effective for fine-tuning (Shen, 2024). Entropy measures the amount of information contained in the generated QA pairs.

In advanced language model evaluation, we use Pair-wise Comparison and Rank Comparison to evaluate the responses of the fine-tuned model on from Students (Zaman et al., 2024) and the questions generated by the LLM based on books(see in Appendix A).

Studies have demonstrated that GPT-4's consistency with human experts as judges reaches 85%, despite limitations such as positional bias (Zheng et al., 2023). In this evaluation, we mitigate positional bias by alternating positions or randomly assigning positions, and we introduce reference-guided evaluation to improve accuracy. This involves having the model generate explanations before making comparisons. Recent research indicates that LLMs can distinguish between utility and relevance and are more effective at identifying evidence helpful for answering questions when using utility judgments (Zhang et al., 2024). During the evaluation process, we input the original text passage corresponding to the test question into the large model as a reference to enhance the quality of its utility judgments.

In Pair-wise Comparison, we use GPT-4 and Claude3-Sonnet as the models. These models score two answers to the same question, and only when one answer wins in both models' comparisons is it declared the winner. If the comparison results of the two models are inconsistent, it is considered a tie. In Rank Comparison, we use GPT-4 to score each model's output based on specific metrics and scales(see prompt in Appendix C.2) We then convert their scores into rankings for each model and each evaluator and take the average (Sottana et al., 2023).

Evaluation Datasets A recent study examined the generalization ability of fine-tuned large language models and designed an evaluation method

Figure 2: The visualization of sentence embeddings from different datasets in a t-SNE two-dimensional space. The blue and red points represent the In-distribution and Out-of-distribution datasets, respectively, used to evaluate the model's performance in different scenarios. The gray points indicate the distribution of the fine-tuned data "book2qa_sft", showing a significant difference from the Out-of-distribution dataset while having a high similarity to the In-distribution dataset.

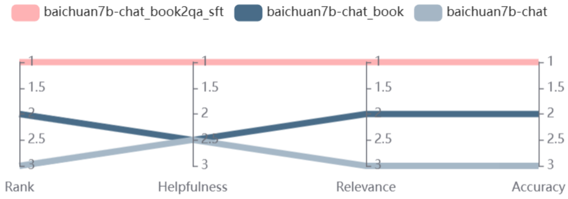

Figure 3: Compare three models on the OOD evaluation set. It shows that the model fine-tuned with book2qa_sft excels in three metric, highlighting its effective fine-tuning and strong generalization ability across different topics. In contrast, the model fine-tuned with book data in an unsupervised manner performs much worse on these metrics.If two models are ranked the same, the results are shown as an average between the lower and upper bounds (e.g., two models both ranked second are shown as 2.5).

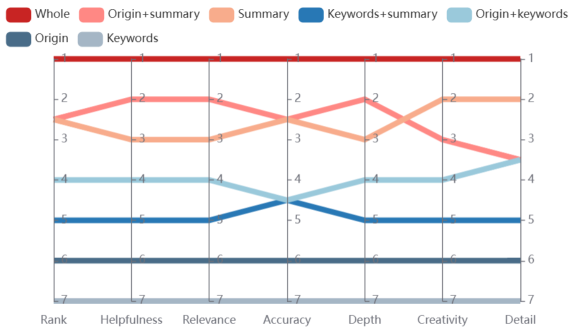

Figure 4: In the GPT-4-based Rank Comparison, models fine-tuned with different data subsets were compared on the ID evaluation set. The model fine-tuned on the whole dataset excelled in all aspects, indicating superior performance on these metrics. This highlights the benefits of integrating multiple information sources original text, keywords, and summaries into the fine-tuning process.

that includes In-domain Datasets and Out-of-domain Datasets for the same tasks (Sottana et al., 2023). Due to budget and time constraints, we selected 240 questions of varying complexity from the DCSC(40 questions per difficulty level across six levels) as the Out-of-distribution evaluation set. Inspired by domain-specific evaluation sets in the literature (Yang et al., 2024), we generated 240 questions based on book content (80 difficult, 80 moderate, and 80 simple) as the In-distribution evaluation set. The Out-of-distribution evaluation set assesses the model's extrapolation and generalization performance in real educational scenarios, while the In-distribution evaluation set evaluates the impact of different training data generation methods in the Book2QA process on the model's response quality.

The spatial distribution of the Out-of-distribution and In-distribution evaluation sets, visualized using tSNE, is shown in Figure 2. The gray data points represent the spatial distribution of book2qa_sft, indicating a significant difference from the Out-of-distribution evaluation set and a closer similarity to the In-distribution evaluation set(see the detailed data information in the appendix A).

5.2 Model Generalization Evaluation

To assess the generalization ability of the model fine-tuned using the book2qa_sft dataset, we established two key baselines: the base model and the unsupervised fine-tuning model (fine-tuning the student model using book data). We used GPT-4 for Rank Comparison of the model's answers on the Out-of-distribution dataset.

As seen in Figure 3, the model fine-tuned with the book2qa_sft dataset outperforms the base model on the Out-of-distribution dataset(detailed data can be found in Appendix D.1), indicating that the fine-tuned model performs well in real scenarios. Training with the Book2QA dataset does not reduce the QA model's generalization ability to other courses, this result demonstrates the effectiveness of the Book2QA method.

Specifically, the fine-tuned Book2QA model shows significant improvements in helpfulness, relevance, and accuracy, demonstrating its advan-

	BF1	Length	Entropy	IFD	ROUGE-1
Origin	**0.62**	**62.51**	4.61	0.67	0.14
Keywords	0.61	55.84	4.50	0.73	0.16
Summary	0.61	62.45	4.53	0.69	0.15
Origin+ keywords	0.61	53.84	4.46	0.75	0.16
Origin+ summary	0.61	58.88	4.51	0.72	0.14
Keywords+ summary	0.60	53.23	4.42	0.75	0.15
Whole	**0.62**	60.30	**4.77**	**0.77**	**0.17**

Table 1: The table shows ablation experiment results for question generation methods using BF1, Length, Entropy, IFD, and ROUGE metrics. The comprehensive dataset, using all information sources, scored highest in BF1, informativeness, IFD, and ROUGE, despite generating slightly shorter questions, proving its effectiveness in producing high-quality questions.

	BF1	r-IFD	Length	Entropy	ROUGE-1
Baichuan13b-chat_sft	0.68	0.12	569.89	5.88	0.33
Qwen7b-chat_sft	0.68	0.11	597.44	**6.20**	0.31
Internlm7b-chat_sft	**0.69**	0.09	**679.60**	6.06	0.32
ChatGPT_sft	0.68	0.10	294.51	5.73	0.32
Book2qa_sft	**0.69**	**0.06**	622.55	6.09	**0.34**

Table 2: A comparative evaluation of five models fine-tuned for answer generation in QA systems is provided, showing scores on BF1, r-IFD, Length, Entropy, and ROUGE. The table highlights the effectiveness of the Book2QA method, which integrates answers from multiple models and achieves the highest scores in BF1, r-IFD, and ROUGE.

tages in generating high-quality, highly relevant answers. In contrast, the unsupervised fine-tuning model scores lower on these metrics, indicating that merely relying on book data for fine-tuning is insufficient to enhance the model's generalization ability across different domains. These results suggest that the Book2QA method not only generates high-quality QA data within specific domains but also exhibits good generalization performance in cross-domain evaluations, providing strong support for applications in the education field.

5.3 Evaluation of Question Quality Through Ablation Experiments

In this section, we conducted extensive ablation experiments to verify the effectiveness of our question generation method. The experiments used fine-tuned datasets generated in seven ways: origin (questions generated using only the original passage information), keywords (questions generated using the original passage and keyword information), summary (questions generated using the original passage and passage summary), origin+keywords (a combination of origin and keywords), origin+summary, keywords+summary, and whole (questions generated using all three information sources). After generating the questions, clustering and IFD filtering were applied to ensure dataset consistency, and qwen7b-chat was used to generate the answers.

Firstly, we conducted an automatic evaluation of these seven datasets, with the results shown in Table 1. We found that the whole dataset achieved the best scores in BF1, question informativeness, IFD, and ROUGE, although it slightly lagged behind the origin and summary datasets in generating question length. Subsequently, we used GPT-4 for Rank Comparison of the responses from the seven fine-tuned models on the in-domain evaluation datasets, as shown in Figure 4. The model fine-tuned on the whole dataset outperformed others in helpfulness, relevance, accuracy, depth, creativity, and detail (detailed data can be found in Appendix D.2). This indicates that integrating multiple information sources can generate higher quality, more diverse, and in-depth questions. Additionally, these results show that the comprehensive use of different prompting strategies can significantly enhance the quality and usability of generated Q&A data, providing strong support for intelligent Q&A systems in the education field.

5.4 Answer Quality Evaluation

This section aims to demonstrate through detailed comparative experiments that Book2QA can effectively integrate outputs from multiple models. Table 2 shows the scores of five fine-tuned datasets in automatic evaluations. The questions in each QA dataset are the same, but different models are used to generate the corresponding answers. For example, baichuan13b-chat_sft contains answers generated by baichuan13b-chat, chatgpt_sft contains answers generated by ChatGPT, and book2qa_sft contains answers from a dataset that merges answers from three models.

The automatic evaluation results indicate that book2qa_sft performs best in BF1, r-IFD, and ROUGE scores, highlighting its superior performance in generating accurate and relevant answers. The answers from Qwen7b-chat_sft are notably rich in information, while Internlm7b-chat_sft consistently produces the longest responses, demonstrating its capability to provide detailed and com-

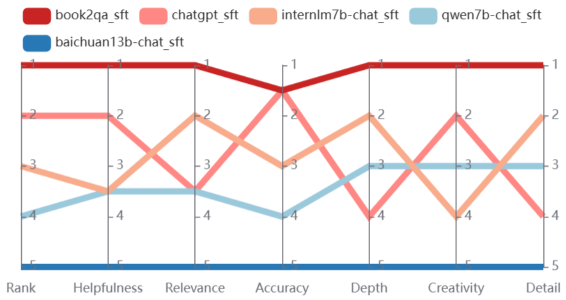

Figure 5: The performance of five models fine-tuned on different datasets was compared using in-domain data on GPT-4. The results highlight that the model fine-tuned with the Book2QA dataset consistently ranked highest across most metrics, demonstrating the superior quality and comprehensiveness of answers generated by the Book2QA method.

prehensive answers.

Next, we performed supervised fine-tuning using the aforementioned five datasets on baichuan7b-chat and used GPT-4 for Rank Comparison of the answers from the five models on an in-domain evaluation dataset. As shown in Figure 5, the Book2QA dataset ranks highest across various metrics (detailed data can be found in Appendix D.3), indicating that integrating answers from different models can produce higher quality answers. This not only improves the richness and diversity of the answers but also enhances the model's adaptability and generalization performance in different contexts.

5.5 Comparative Experiments with SOTA Models

To conduct pairwise comparisons, we fine-tuned the Baichuan7b-chat model using the book2qa_sft, ChatGPTqa, and Qwen-Maxqa datasets. Among these, ChatGPTqa and Qwen-Maxqa are question-answering datasets generated using the data generation process proposed in this paper by the ChatGPT and Qwen-Max models, respectively. Since we did not use IFD and r-IFD scores for filtering, each paragraph generated 30 question-answer pairs, maintaining consistency with the data volume of the book2qa_sft dataset. ChatGPT and Qwen-Max are currently powerful large-scale language models with hundreds of billions of parameters, in English and Chinese respectively.

For the fine-tuned models, we used GPT-4 and Claude3-Sonnet as evaluators to assess responses on the in-domain evaluation datasets. For each instruction, we compared responses using a "win-tie-loss" metric. Specifically, when the evaluation

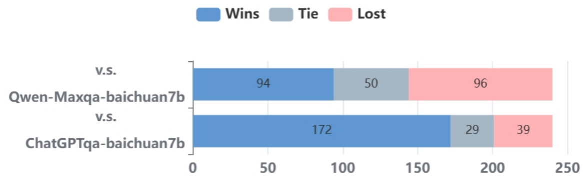

Figure 6: The "win-tie-loss" evaluation results show that our fine-tuned model performed well compared to ChatGPTqa-baichuan7b and was competitive with Qwen-Maxqa-baichuan7b, using GPT-4 and Claude3-Sonnet as evaluators. This indicates that our model, which integrates data from multiple smaller models, can rival or even surpass the capabilities of larger, more complex language models.

results of GPT-4 and Claude3-Sonnet were consistent, we judged it as a win or loss; when the evaluation results were inconsistent, we judged it as a tie. The evaluation results are shown in Figure 6. It is noteworthy that our model significantly outperformed the ChatGPTqa fine-tuned model(ChatGPTqa-baichuan7b) and was slightly inferior to the Qwen-Maxqa fine-tuned model(Qwen-Maxqa-baichuan7b). This demonstrates that the method proposed in this paper, combining data generated by multiple small models, can rival or even surpass the data generated by large-scale language models with hundreds of billions of parameters (detailed data can be found in Appendix D.4).

6 Conclusion

This study introduces a novel method for generating question-answering data from textbooks to aid in fine-tuning educational QA robots across various teaching domains with scarce data. Through both automatic evaluation and assessment by advanced large language models, we demonstrated the effectiveness of the Book2QA approach in integrating outputs from multiple mid-sized models to generate high-quality QA data. Compared to existing large language models, Book2QA not only reduces costs but also performs excellently across various evaluation metrics, providing a robust solution for intelligent QA systems in the educational field. However, upon closer human inspection, we observed several issues with the synthetic data, such as hallucination phenomena. Despite these problems, experimental results indicate that the generated data can be used for further fine-tuning of QA robots, performing well in real-world scenarios and exhibiting a certain degree of generalization capability.

Limitations

Our work relies on the availability of textbook data and is confined to specific teaching scenarios and subjects. Although we have demonstrated that the fine-tuned models possess some degree of generalization capability, these observations are limited to the QA tasks that were considered in the experiments. Future work could focus on other tasks and scenarios.

While achieving the highest scores is not the ultimate goal of education, information-rich answers do reflect students' learning potential. However, too much information may increase the cognitive load on students, making it difficult to digest (Mayer and Moreno, 2003). Therefore, future research needs to carefully balance the amount of information included in dialogues. Similarly, educators can set reasonable target metrics and their combinations as needed to better optimize the fine-tuning performance for educational use cases.

Despite our attempts to incorporate various aspects of evaluation in this work, it is still not possible to cover all aspects of educational assessment. Additionally, educational QA evaluation is highly subjective, challenging to evaluate, and lacks domain-specific evaluation datasets. This paper has made some attempts, such as collecting real-world QA data, generating in-domain evaluation sets, and introducing evaluations by large language models.

Ethics Statement

This study strictly adhered to the ethical guidelines for the application of artificial intelligence in the field of education. During the development and evaluation of the Book2QA method, we used content extracted from textbooks, ensuring that all data were anonymous and did not involve any personally identifiable information. We did not use any real student data that involved privacy risks. We disclosed the methods of data generation and model training, ensuring the transparency of the research process. Furthermore, we will open-source our code and part of the generated data on GitHub. We acknowledge the limitations of our research and have explicitly pointed out the possible hallucination phenomena in the generated data in our paper. We commit to continuing to improve the model in future work to mitigate these issues.

References

Lorin W. Anderson, David R. Krathwohl, and Benjamin Samuel Bloom. 2000. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, George Louis Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024a. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024b. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024c. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning.

Suichan Li, Dongdong Chen, Yinpeng Chen, Lu Yuan, Lei Zhang, Qi Chu, Bin Liu, and Nenghai Yu. 2021. Unsupervised finetuning.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiaoding Lu, Zongyi Liu, Adian Liusie, Vyas Raina, Vineet Mudupalli, Yuwen Zhang, and William Beauchamp. 2024. Blending is all you need: Cheaper, better alternative to trillion-parameters llm.

Richard E. Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38:43 – 52.

Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Ming Shen. 2024. Rethinking data selection for supervised fine-tuning.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks.

Md Arafat Sultan, Shubham Chandel, Ramón Fernández Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for qa. In *Annual Meeting of the Association for Computational Linguistics*.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. Book2dial: Generating teacher-student interactions from textbooks for cost-effective development of educational chatbots.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024b. Large language models for education: A survey and outlook.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias.

Khandoker Ashik Uz Zaman, Ashraful Islam, Yusuf Mahbubul Islam, and Md Abu Sayed. 2024. Dataset of computer science course queries from students: Categorized and scored according to bloom's taxonomy. *Data in Brief*, 53:110109.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are large language models good at utility judgments?

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Datasets

We briefly describe the datasets used in our method and experiments. Table 3 shows the statistics for the textbooks, real Q&A data and test sets in the evaluation datasets. The following includes detailed information about the datasets:

- **Data and Programming** delves deeply into the basic concepts of data, principles of digitization and encoding, and the representation and processing of data in computer systems. The book also introduces big data technology, data collection and processing methods, as well as data analysis and visualization techniques, providing readers with a comprehensive perspective on the application and development trends of information technology in the field of data science.

- **Information Systems and Digital Society** focuses on discussing the components and functions of information systems and their extensive application in modern society. The book analyzes the development history of information technology, explores the characteristics of the digital society and the rise of the digital industry, as well as the application experience of artificial intelligence technology in

Dataset	Chapters	Words	Examples
Data and Programming	14	81090	-
Information Systems&Digital Society	14	129610	-
DCSCQ	-	127556	8811
OOD evaluation set	-	12152	240
ID evaluation set	-	5154	240

Table 3: Performed statistical analysis on different datasets.

real life, revealing how information technology promotes social progress and improves the quality of people's lives.

- **DCSCQ** is a dataset of real student questions related to computer science courses, including questions verified by keywords and evaluated for complexity using Bloom's taxonomy, covering subjects not included in the textbooks used in this study, such as data structures. we selected 240 questions of varying complexity from the DCSC(40 questions per difficulty level across six levels) as the **Out-of-distribution evaluation set**(Detailed data examples are in the appendix E.2).

- **In-distribution evaluation set** utilizes the same prompting process as Book2QA, leveraging cloude3-sonnet and based on "Data and Programming" and "Information Systems and Digital Society," it has generated 240 test questions and divided them into three different levels of difficulty, including easy, moderate, and difficult. The purpose is to evaluate the performance of the fine-tuned model on test questions of varying difficulty(Detailed data examples are in the appendix E.2).

B Data Preprocessing Details

We used ChatGPT to generate summaries and keywords for each paragraph in the books(Detailed data examples are in the appendix E.3). We ultimately extracted 198 paragraphs from two books and divided them into 14 clustering clusters, with the number of summaries in each cluster shown in Figure 7. The number of keywords was set to three keywords per paragraph. The generated summaries and keywords will be input into the large language model in different ways during the data generation phase, as detailed in Appendix C.

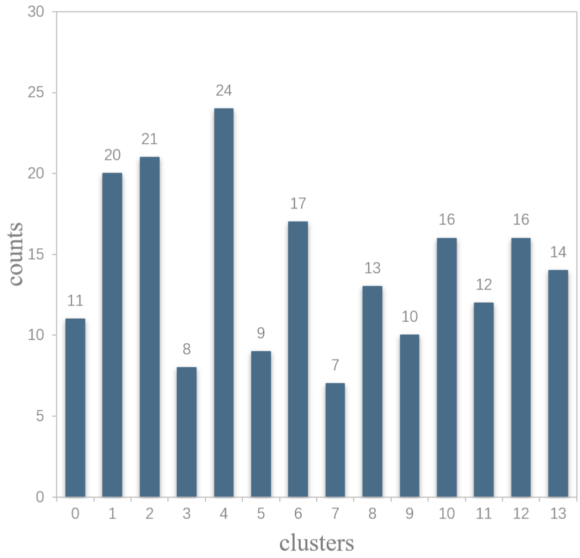

Figure 7: The distribution of paragraph summaries extracted from two books across 14 clusters is presented. The histogram illustrates the number of summaries in each cluster, where the largest cluster contains 24 summaries and the smallest contains 7.

C prompts for data generation and evaluation

C.1 prompts for data generation

We developed three prompting strategies based on Bloom's Taxonomy to provide varying levels of factual information to large language models: the first strategy uses only the original text passage; the second strategy adds paragraph summaries from different groups to the original text; the third strategy includes a keyword in the original text passage. The prompts based on Bloom's Taxonomy and the three strategies are shown in Figure 11.

C.2 prompts for evaluation

We developed two prompting strategies to evaluate advanced large language models through rank comparison and pair comparison. For the rank comparison, GPT-4 was prompted to rate multiple candidate answers to the same question using a Likert five-point scale, assessing helpfulness, relevance, accuracy, depth, creativity, and detail (Section 4.2: Model Generalization Evaluation focused only on the first three aspects). For the pair comparison, GPT-4 and Claude3-Sonnet were used as evaluators to assess responses on in-domain evaluation datasets. Each instruction's responses were com-

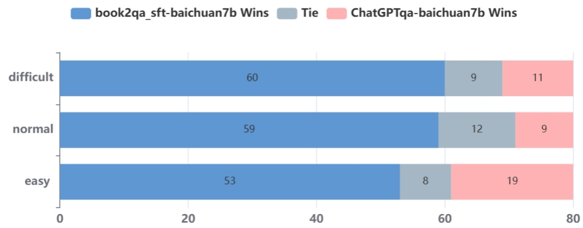

Figure 8: The results of our model and the Pair-wise Comparison using the ChatGPTqa fine-tuning model at different difficulty levels.

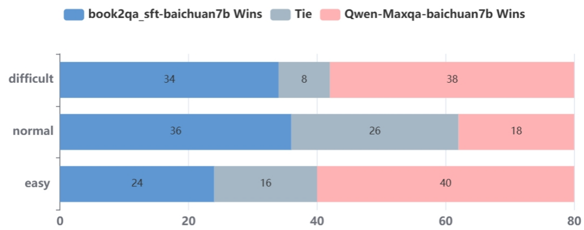

Figure 9: The results of our model and the Pair-wise Comparison using the Qwen-Maxqa fine-tuning model at different difficulty levels.

pared using a "win-tie-loss" metric. Consistent evaluation results from GPT-4 and Claude3-Sonnet were judged as a win or loss, while inconsistent results were judged as a tie. Specific evaluation prompts can be found in Figure 12.

D Detailed Experimental Data

D.1 Model Generalization Evaluation Data

Firstly, we used GPT-4 to score the responses of three models on the OOD evaluation set using a 1-5 Likert scale, where 1 indicates the worst performance and 5 indicates the best performance. We scored the responses on helpfulness, relevance, and accuracy, and provided a total score for each response. By scoring the answers of 240 evaluation questions(Detailed result examples are in the appendix E.4), we obtained detailed experimental data shown in Table 4, which presents the average scores and variances of different models across the six levels of questions based on Bloom's taxonomy. Subsequently, we converted these scores into rankings, as shown in Table 5.

D.2 Ablation Experiments Data

In the ablation experiment, we used models fine-tuned on data generated based on different amounts of information to score answers on the ID evaluation set using a Likert scale from 1 to 5. We scored the answers based on helpfulness, relevance, accuracy, depth, creativity, and thoroughness, and

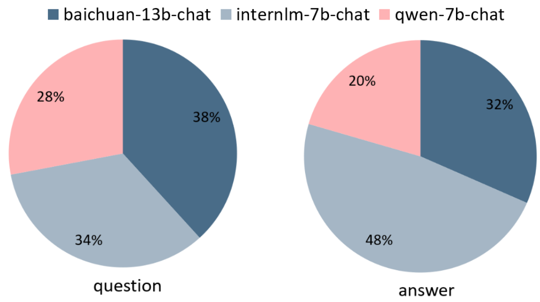

Figure 10: Two pie charts comparing the contributions of three AI models: "baichuan-13b-chat," "interlim-7b-chat," and "qwen-7b-chat" in generating questions and answers. "baichuan-13b-chat" contributes the most in the "Questions" pie chart, while "qwen-7b-chat" leads in the "Answers" pie chart.

provided an overall score for each answer. By scoring 240 evaluation questions, we obtained detailed experimental data, as shown in Table 6, which presents the average scores and variances of various models across different difficulty levels. Subsequently, we converted these scores into rankings, as shown in Table 7.

D.3 Comparative Experiments Data

In the comparative experiment, we used models fine-tuned on answers generated differently to score responses on the ID evaluation set using a Likert scale from 1 to 5. It is important to note that in the QA fine-tuning data used in this section, the questions are the same, but the answers are provided by different models. We scored the answers based on helpfulness, relevance, accuracy, depth, creativity, and thoroughness, and provided an overall score for each answer. By scoring 240 evaluation questions, we obtained detailed experimental data, as shown in Table 8, which presents the average scores and variances of various models across different difficulty levels. Subsequently, we converted these scores into rankings, as shown in Table 9.

D.4 Pair-wise Comparison Experiments Data

In Pair-wise Comparison, detailed experimental data is shown in Figures 8 and 9. These figures respectively present the Pair-wise Comparison of models fine-tuned using book2qa_sft with those fine-tuned using QA data generated by ChatGPT and Qwen-Max on the ID evaluation set.

score segments	Model	avg. s	std. s	avg. h	std. h	avg. r	std. r	avg. a	std. a
0-20	baichuan7b-chat_book2qa_sft	4.47	0.44	4.45	0.44	4.50	0.44	4.48	0.45
	baichuan7b-chat_book	4.03	0.26	3.99	0.28	4.10	0.34	4.04	0.28
	baichuan7b-chat	3.89	0.57	4.00	0.61	3.94	0.61	3.88	0.63
21-40	baichuan7b-chat_book2qa_sft	4.43	0.27	4.44	0.27	4.41	0.25	4.45	0.30
	baichuan7b-chat_book	4.23	0.38	4.20	0.37	4.20	0.42	4.16	0.51
	baichuan7b-chat	4.10	0.49	4.32	0.48	4.03	0.50	4.02	0.58
41-60	baichuan7b-chat_book2qa_sft	4.50	0.47	4.49	0.46	4.48	0.46	4.55	0.50
	baichuan7b-chat_book	4.14	0.21	4.14	0.21	4.15	0.21	4.16	0.26
	baichuan7b-chat	3.98	0.47	4.10	0.49	3.96	0.47	3.97	0.48
61-80	baichuan7b-chat_book2qa_sft	4.48	0.18	4.49	0.16	4.45	0.21	4.50	0.17
	baichuan7b-chat_book	4.22	0.16	4.25	0.16	4.18	0.17	4.24	0.17
	baichuan7b-chat	4.30	0.11	4.34	0.13	4.27	0.12	4.38	0.30
81-100	baichuan7b-chat_book2qa_sft	4.42	0.18	4.42	0.17	4.41	0.18	4.45	0.23
	baichuan7b-chat_book	4.29	0.21	4.34	0.18	4.27	0.17	4.29	0.29
	baichuan7b-chat	4.16	0.45	4.20	0.49	4.15	0.39	4.14	0.67
100+	baichuan7b-chat_book2qa_sft	4.38	0.23	4.36	0.25	4.37	0.23	4.40	0.22
	baichuan7b-chat_book	4.07	0.23	4.06	0.24	4.07	0.22	4.07	0.22
	baichuan7b-chat	3.95	0.37	4.06	0.36	3.95	0.37	3.95	0.37
Total	baichuan7b-chat_book2qa_sft	4.44	0.29	4.44	0.29	4.44	0.30	4.47	0.31
	baichuan7b-chat_book	4.16	0.25	4.16	0.26	4.16	0.26	4.16	0.30
	baichuan7b-chat	4.06	0.43	4.17	0.44	4.05	0.42	4.05	0.53

Table 4: Presents the average scores and variances of different models across the six levels of questions based on Bloom's taxonomy.In the table, s represents the total score, h represents helpfulness, r represents relevance, and a represents accuracy.

score segments	Model	avg. s	avg. h	avg. r	avg. a
0-20	baichuan7b-chat_book2qa_sft	1	1	1	1
	baichuan7b-chat_book	2	3	2	2
	baichuan7b-chat	3	2	3	3
21-40	baichuan7b-chat_book2qa_sft	1	1	1	1
	baichuan7b-chat_book	2	2	2	2
	baichuan7b-chat	3	3	3	3
41-60	baichuan7b-chat_book2qa_sft	1	1	1	1
	baichuan7b-chat_book	2	2	2	2
	baichuan7b-chat	3	3	3	3
61-80	baichuan7b-chat_book2qa_sft	2	2	2	2
	baichuan7b-chat_book	3	3	3	3
	baichuan7b-chat	1	1	1	1
81-100	baichuan7b-chat_book2qa_sft	2	2	2	1
	baichuan7b-chat_book	1	1	1	2
	baichuan7b-chat	3	3	3	3
100+	baichuan7b-chat_book2qa_sft	1	1	1	1
	baichuan7b-chat_book	2.5	2.5	2	2.5
	baichuan7b-chat	2.5	2.5	2	2.5
Total	baichuan7b-chat_book2qa_sft	1	1	1	1
	baichuan7b-chat_book	2	2.5	2	2
	baichuan7b-chat	3	2.5	3	3

Table 5: Convert scores to rankings,if two models are ranked the same, the results are shown as an average between the lower and upper bounds (e.g., two models both ranked second are shown as 2.5)

Difficulty	Model	avg. s	std. s	avg. h	std. h	avg. r	std. r	avg. a	std. a	avg. d	std. d	avg. c	std. c	avg. t	std. t
easy	origin	4.20	0.63	4.11	0.64	4.24	0.67	4.19	0.65	4.18	0.66	4.01	0.59	4.16	0.63
	keywords	4.08	0.56	3.91	0.58	4.08	0.58	4.09	0.61	3.95	0.57	3.96	0.52	3.97	0.53
	summary	4.36	0.63	4.29	0.70	4.43	0.66	4.31	0.59	4.26	0.62	4.24	0.66	4.27	0.61
	origin+keywords	4.35	0.54	4.31	0.55	4.37	0.56	4.33	0.56	4.31	0.56	4.21	0.51	4.31	0.55
	origin+summary	4.44	0.62	4.46	0.67	4.52	0.70	4.39	0.63	4.42	0.62	4.33	0.59	4.34	0.59
	keywords+summary	4.25	0.54	4.15	0.57	4.22	0.57	4.25	0.59	4.16	0.53	4.13	0.54	4.20	0.55
	whole	4.71	0.61	4.78	0.64	4.79	0.64	4.77	0.64	4.71	0.64	4.51	0.57	4.73	0.61
normal	origin	4.06	0.23	4.00	0.19	4.03	0.29	3.99	0.55	3.82	0.46	3.40	0.46	3.99	0.45
	keywords	3.91	0.34	3.86	0.33	3.86	0.51	3.78	0.45	3.65	0.32	3.31	0.53	3.88	0.37
	summary	4.19	0.41	4.19	0.47	4.19	0.53	4.10	0.64	3.89	0.54	3.60	0.48	4.27	0.64
	origin+keywords	4.13	0.33	4.15	0.37	4.17	0.44	3.92	0.46	3.89	0.49	3.58	0.42	4.25	0.52
	origin+summary	4.05	0.40	4.08	0.52	4.07	0.39	3.95	0.61	3.85	0.61	3.37	0.59	4.11	0.78
	keywords+summary	4.11	0.24	4.11	0.35	4.15	0.35	3.90	0.27	3.83	0.33	3.55	0.38	4.15	0.47
	whole	4.39	0.50	4.54	0.59	4.43	0.62	4.26	0.70	4.13	0.54	3.70	0.67	4.53	0.59
difficult	origin	4.14	0.08	4.07	0.18	4.21	0.17	4.02	0.05	4.05	0.42	3.89	0.25	4.46	0.19
	keywords	4.06	0.09	4.03	0.10	4.03	0.17	3.99	0.10	3.96	0.47	3.84	0.19	4.28	0.24
	summary	4.58	0.04	4.82	0.12	4.67	0.13	4.57	0.17	4.38	0.25	4.15	0.08	4.70	0.17
	origin+keywords	4.45	0.06	4.65	0.15	4.62	0.20	4.36	0.19	4.25	0.26	4.05	0.13	4.54	0.16
	origin+summary	4.61	0.09	4.79	0.15	4.75	0.14	4.66	0.16	4.48	0.24	4.20	0.18	4.66	0.20
	keywords+summary	4.50	0.04	4.72	0.15	4.65	0.14	4.48	0.25	4.31	0.17	4.08	0.05	4.64	0.15
	whole	4.82	0.10	4.91	0.10	4.92	0.07	4.85	0.15	4.79	0.24	4.54	0.25	4.94	0.05
total	origin	4.13	0.32	4.06	0.34	4.16	0.38	4.06	0.43	4.01	0.53	3.75	0.51	4.20	0.46
	keywords	4.01	0.33	3.93	0.34	3.99	0.43	3.95	0.41	3.85	0.47	3.69	0.50	4.04	0.41
	summary	4.37	0.39	4.42	0.51	4.42	0.49	4.32	0.51	4.17	0.52	3.99	0.50	4.41	0.52
	origin+keywords	4.30	0.33	4.36	0.40	4.38	0.44	4.20	0.45	4.15	0.48	3.94	0.43	4.36	0.43
	origin+summary	4.37	0.43	4.43	0.54	4.44	0.49	4.32	0.56	4.24	0.58	3.96	0.65	4.36	0.58
	keywords+summary	4.28	0.31	4.32	0.44	4.33	0.41	4.20	0.43	4.09	0.39	3.91	0.40	4.32	0.44
	whole	4.63	0.44	4.74	0.48	4.71	0.50	4.62	0.58	4.54	0.57	4.24	0.66	4.72	0.46

Table 6: The detailed data of the ablation experiment shows the average scores and variances of different models based on three difficulty levels. In the table, s represents the total score, h represents helpfulness, r represents relevance, a represents accuracy, d represents depth, c represents creativity, and t represents thoroughness.

Difficulty	Model	avg. s	avg. h	avg. r	avg. a	avg. d	avg. c	avg. t
easy	Origin	6	6	5	6	5	6	6
	Keywords	7	7	7	7	7	7	7
	Summary	3	4	3	4	4	3	4
	Origin+Keywords	4	3	4	3	3	4	3
	Origin+Summary	2	2	2	2	2	2	2
	Keywords+Summary	5	5	6	5	6	5	5
	Whole	1	1	1	1	1	1	1
normal	Origin	5	6	6	3	6	5	6
	Keywords	7	7	7	7	7	7	7
	Summary	2	2	2	2	2.5	2	2
	Origin+Keywords	3	3	3	5	2.5	3	3
	Origin+Summary	6	5	5	4	4	6	5
	Keywords+Summary	4	4	4	6	5	4	4
	Whole	1	1	1	1	1	1	1
difficult	Origin	6	6	6	6	6	6	6
	Keywords	7	7	7	7	7	7	7
	Summary	3	2	3	3	3	3	2
	Origin+Keywords	5	5	5	5	5	5	5
	Origin+Summary	2	3	2	2	2	2	3
	Keywords+Summary	4	4	4	4	4	4	4
	Whole	1	1	1	1	1	1	1
total	Origin	6	6	6	6	6	6	6
	Keywords	7	7	7	7	7	7	7
	Summary	2.5	3	3	2.5	3	2	2
	Origin+Keywords	4	4	4	4.5	4	4	3.5
	Origin+Summary	2.5	2	2	2.5	2	3	3.5
	Keywords+Summary	5	5	5	4.5	5	5	5
	Whole	1	1	1	1	1	1	1

Table 7: The ablation experiment score data has been converted to rankings.

Difficulty	Model	avg. s	std. s	avg. h	std. h	avg. r	std. r	avg. a	std. a	avg. d	std. d	avg. c	std. c	avg. t	std. t
easy	baichuan13b-chat_sft	3.75	0.88	3.74	0.86	3.85	0.77	3.75	0.88	3.40	1.30	2.83	0.75	3.58	1.22
	qwen7b-chat_sft	4.62	0.62	4.62	0.62	4.69	0.60	4.62	0.62	4.32	0.63	3.75	0.38	4.57	0.63
	internlm7b-chat_sft	4.67	0.29	4.67	0.29	4.74	0.15	4.66	0.32	4.59	0.47	3.67	0.34	4.67	0.34
	chatgpt_sft	4.48	0.19	4.48	0.19	4.59	0.09	4.54	0.05	4.02	0.55	3.57	0.60	4.21	0.40
	book2qa_sft	4.47	0.78	4.50	0.75	4.60	0.66	4.47	0.78	4.14	1.18	3.61	0.93	4.38	0.96
normal	baichuan13b-chat_sft	3.79	0.48	3.94	0.70	3.83	0.70	3.85	0.55	3.39	0.73	3.41	0.57	4.05	0.54
	qwen7b-chat_sft	4.52	0.08	4.76	0.14	4.69	0.22	4.59	0.25	4.28	0.24	4.09	0.12	4.59	0.20
	internlm7b-chat_sft	4.60	0.39	4.82	0.43	4.76	0.48	4.70	0.46	4.52	0.61	4.12	0.47	4.71	0.41
	chatgpt_sft	4.80	0.09	4.93	0.02	4.77	0.22	4.92	0.05	4.34	0.47	4.33	0.47	4.64	0.26
	book2qa_sft	4.82	0.05	4.96	0.03	4.94	0.04	4.83	0.11	4.82	0.14	4.49	0.25	4.95	0.04
difficult	baichuan13b-chat_sft	3.48	0.61	3.55	0.66	3.64	0.83	3.31	0.72	3.10	0.76	3.04	0.69	3.86	0.62
	qwen7b-chat_sft	4.41	0.12	4.62	0.27	4.68	0.24	4.33	0.22	4.21	0.32	4.01	0.20	4.52	0.28
	internlm7b-chat_sft	4.40	0.17	4.54	0.30	4.66	0.25	4.35	0.32	4.31	0.37	3.88	0.35	4.64	0.25
	chatgpt_sft	4.44	0.12	4.65	0.23	4.71	0.20	4.40	0.30	4.15	0.29	3.98	0.21	4.48	0.28
	book2qa_sft	4.56	0.08	4.77	0.17	4.81	0.15	4.52	0.26	4.42	0.27	4.01	0.25	4.81	0.18
total	baichuan13b-chat_sft	3.67	0.68	3.74	0.76	3.77	0.78	3.62	0.77	3.29	0.94	3.09	0.72	3.83	0.82
	qwen7b-chat_sft	4.51	0.27	4.67	0.34	4.69	0.35	4.51	0.37	4.27	0.40	3.95	0.25	4.56	0.37
	internlm7b-chat_sft	4.55	0.29	4.67	0.35	4.72	0.30	4.56	0.39	4.46	0.50	3.89	0.42	4.67	0.33
	chatgpt_sft	4.57	0.16	4.68	0.19	4.69	0.18	4.61	0.19	4.17	0.45	3.96	0.51	4.45	0.34
	book2qa_sft	4.62	0.31	4.74	0.34	4.78	0.30	4.61	0.40	4.47	0.59	4.04	0.59	4.72	0.44

Table 8: The detailed data of the comparative experiment shows the average scores and variances of different models based on three difficulty levels. In the table, s represents the total score, h represents helpfulness, r represents relevance, a represents accuracy, d represents depth, c represents creativity, and t represents thoroughness.

Difficulty	Model	avg. s	avg. h	avg. r	avg. a	avg. d	avg. c	avg. t
easy	baichuan13b-chat_sft	5	5	5	5	5	5	5
	qwen7b-chat_sft	2	2	2	2	2	1	2
	internlm7b-chat_sft	1	1	1	1	1	2	1
	chatgpt_sft	3	4	4	3	4	4	4
	book2qa_sft	4	3	3	4	3	3	3
normal	baichuan13b-chat_sft	5	5	5	5	5	5	5
	qwen7b-chat_sft	4	4	4	4	4	4	4
	internlm7b-chat_sft	3	3	3	3	2	3	2
	chatgpt_sft	2	2	2	1	3	2	3
	book2qa_sft	1	1	1	2	1	1	1
difficult	baichuan13b-chat_sft	5	5	5	5	5	5	5
	qwen7b-chat_sft	3	3	3	4	3	1.5	3
	internlm7b-chat_sft	4	4	4	3	2	4	2
	chatgpt_sft	2	2	2	2	4	3	4
	book2qa_sft	1	1	1	1	1	1.5	1
total	baichuan13b-chat_sft	5	5	5	5	5	5	5
	qwen7b-chat_sft	4	3.5	3.5	4	3	3	3
	internlm7b-chat_sft	3	3.5	2	3	2	4	2
	chatgpt_sft	2	2	3.5	1.5	4	2	4
	book2qa_sft	1	1	1	1.5	1	1	1

Table 9: The comparative experiment score data has been converted to rankings.

E Data Example

E.1 Example of Generated QA Data

Figure 13 shows examples of question-answer pairs in the book2qa_sft dataset generated using the BOOK2QA method with the number of question clusters k set to 30. The distribution of the sources of questions and answers is shown in Figure 10.

E.2 Example of Evaluation Data

Figure 14 shows examples of questions of varying difficulty from the ID evaluation dataset, and Figure 15 shows examples of questions from the OOD evaluation dataset categorized by Bloom's taxonomy scoring.

E.3 Example of Generated Datasets Summary and Keyword

We provide the summary and keywords of our generated dataset in Figure 16.In this paper, we set the number of generated keywords n to 3.

E.4 Example of Model Response in Evaluation Experiment

In Figure 17, we provide partial examples of responses from different models on the OOD test set in the Model Generalization Evaluation.

prompt 1

##instruction

你是一个出题人，基于[文档]，请提出一个[问题]，提出的[问题]要以[文档]的要点为基础，并满足以下要求：(You are an examiner. Based on the [document], please propose a [question] that is based on the key points of the [document] and meets the following requirements:)

{{a prompt from Bloom's Taxonomy}}

[问题]要尽可能复杂。(The [question] should be as complex as possible.)

##document

{{paragraph}}

##output format

[问题]：提出的问题([Question]: The proposed question)

prompt 2

##指令

你是一个出题人，基于[文档]，请提出一个[问题]，提出的[问题]要以[文档]的要点为基础，并满足以下要求：(You are an examiner. Based on the [document], please propose a [question] that is based on the key points of the [document] and meets the following requirements:)

{{a prompt from Bloom's Taxonomy}}

[问题]要尽可能复杂。提出的[问题]要关于[主题]:(The [question] should be as complex as possible.The proposed [question] should be about the [topic]:)

{{keywords}}

##document

{{paragraph}}

##output format

[问题]：提出的问题([Question]: The proposed question)

prompt 3

##instruction

你是一个出题人，基于[文档]，请提出一个[问题]，提出的[问题]要以[文档]的要点为基础，当[问题]需要额外信息时再参考[参考摘要],并满足以下要求：(You are an examiner. Based on the [document], please propose a [question] that is based on the key points of the [document] and, when the [question] requires additional information, refer to the [reference summary]. The question should meet the following requirements:)

{{a prompt from Bloom's Taxonomy}}

[问题]要尽可能复杂。(The [question] should be as complex as possible.)

##Reference Summary

与[文档]主题相同的[参考摘要]:(Reference summary related to the [document]'s topic:){{relevant summary}}

与[文档]主题不同的[参考摘要]:(Reference summary unrelated to the [document]'s topic){{irrelevant summary}}

##document

{{paragraph}}

##output format

[问题]：提出的问题([Question]: The proposed question)

Figure 11: The prompts based on Bloom's Taxonomy and the three strategies.

```
prompts based on Bloom's Taxonomy
prompt_list = [
	'提出认知性问题，对[文档]中知识的回忆与确认。用一种非常接近于学生当初遇到的某种观念和现象时的形式，进行提问。提示词：回忆，记忆，识别，列表，定义，陈述，呈现(Propose a cognitive question that recalls and confirms the knowledge in the [document]. Ask the question in a form very similar to the one encountered by students when they first encounter certain concepts and phenomena. Prompt words: recall, memorize, recognize, list, define, state, present)',
	'提出理解性问题，对[文档]中有关概念内容的理解进行提问，[问题]的回答需要根据[文档]进行总结和比较。提示词: 说明，识别，描述，解释，区别，重述，归纳，比较(Propose a comprehension question that asks about the understanding of concepts in the [document]. The answer to the [question] needs to summarize and compare based on the [document]. Prompt words: explain, identify, describe, interpret, distinguish, restate, summarize, compare)',
	'提出分析性问题，问题的回答要利用[文档]中的内容进行推理与解析，详细地阐明基础理论和基本原理。提示词: 分析，检查，实验，组织，对比，比较，辨别，区别(Propose an analytical question. The answer to the question should use the content of the [document] for reasoning and analysis, and explain the basic theories and principles in detail. Prompt words: analyze, examine, experiment, organize, contrast, compare, differentiate, distinguish)',
	'提出评价性问题，问题的回答要根据[文档]作出合适的判断，并提出有说服力的观点。提示词: 评价，估计，评论，鉴定，辩明，辩护，证明，预测，预言，支持(Propose an evaluative question. The answer to the question should make an appropriate judgment based on the [document] and provide a convincing opinion. Prompt words: evaluate, estimate, comment, appraise, justify, prove, predict, forecast, support)',
	'提出应用性问题，问题的回答要利用[文档]和[参考摘要]中的信息解决实际的问题，涉及具有特色的表达，制定合理的计划和可实施的步骤，根据基本材料推出某种规律等活动。提示词: 应用，论证，操作，实践，分类，举例说明，解决(Propose an application question. The answer to the question should use information from the [document] and the [reference summary] to solve practical problems, involving unique expressions, formulating reasonable plans and implementable steps, and deriving certain rules based on basic materials. Prompt words: apply, argue, operate, practice, classify, illustrate, solve)',
	'提出综合性问题，对[文档]的内容进行发散，可以做一些假设，利用[文档]和[参考摘要]的内容进行系统的分析，问题的回答要对事物本质的价值作出有说服力的判断。提示词: 组成，建立，设计，开发，计划，支持，系统化(Propose a synthesis question that diverges from the content of the [document]. You can make some assumptions and use the content of the [document] and [reference summary] for systematic analysis. The answer to the question should make a convincing judgment about the essence of things. Prompt words: compose, create, design, develop, plan, support, systematize)'
]
```

Figure 12: Evaluation prompts for advanced large language models.

Question Type: Cognitive question

Question: 请回忆并描述信息系统中的五个主要部分，以及它们各自的组成和功能。(Please recall and describe the five main components of an information system, along with their respective compositions and functions.)

Output: 信息系统由五个主要部分组成：硬件、软件、网络、数据和用户...(An information system consists of five main components: hardware, software, networks, data, and users...)

Question Type: Comprehension question

Question: 请解释计算机网络结构是如何被划分为两层，并说明这两层分别提供什么服务？(Please explain how the computer network architecture is divided into two layers and describe the services provided by each layer.)

Output: 计算机网络结构被划分为两层：外层为由主机构成的资源子网，内层为由通信设备和通信线路构成的通信子网...(The computer network architecture is divided into two layers: the outer layer is the resource subnet composed of hosts, and the inner layer is the communication subnet composed of communication devices and communication lines....)

Question Type: Analytical question

Question: 请分析并比较高级语言C、C++、Java、C#和Python在语言特点上的异同，并阐述它们在不同应用领域中的优势和局限性。(Please analyze and compare the similarities and differences in language characteristics of high-level languages C, C++, Java, C#, and Python, and explain their advantages and limitations in different application domains.)

Output: 高级语言C、C++、Java、C#和Python在语言特点上有显著的异同：\n\n1. **C/C++**...(High-level languages C, C++, Java, C#, and Python have significant similarities and differences in their language characteristics:\n\n1. **C/C++**...)

Question Type: Evaluative question

Question: 如何评价网络安全法对保障网络安全的重要性，并分析其对社会发展和个人权益保护的影响？(How to evaluate the importance of the Cybersecurity Law in ensuring network security, analyze its impact on social development and the protection of individual rights?)

Output: 网络安全法对保障网络安全的重要性体现在多个方面。首先...(The importance of the Cybersecurity Law in ensuring network security is reflected in multiple aspects. First...)

Question Type: Application question

Question: 如何利用信息的普遍性特征来解决一个实际的社会问题，比如提高城市公共交通的效率？(How can the universality of information be used to solve a practical social problem, such as improving the efficiency of urban public transportation?)

Output: 要利用信息的普遍性特征来解决一个实际的社会问题，比如提高城市公共交通的效率，可以采取以下步骤...(To utilize the universality of information to solve a practical social problem, such as improving the efficiency of urban public transportation, the following steps can be taken...)

question type: synthesis question

Question: 如何通过系统化的方法设计和开发一个能够有效管理和验证信息真伪，同时根据用户需求提供个性化价值的综合信息系统，以支持社会信息道德规范并抵御网络攻击？(How to design and develop a comprehensive information system that can effectively manage and verify the authenticity of information, while providing personalized value based on user needs, in order to support social information ethics and resist cyber attacks through a systematic approach?)

Output: 要设计和开发一个能够有效管理和验证信息真伪，同时根据用户需求提供个性化价值的综合信息系统，以支持社会信息道德规范并抵御网络攻击，可以通过以下系统化的方法...(To design and develop a comprehensive information system that can effectively manage and verify the authenticity of information, while providing personalized value based on user needs, in order to support social information ethics and resist cyber attacks, the following systematic approach can be taken...)

Figure 13: Example generations of QAs in the book2qa_sft dataset based on Bloom's taxonomy.Due to the length of some answers, ellipses (...) are used for omission.

Question Classification: easy

Question1: CPU由哪两大部分组成?(What are the two main components of the CPU?)

Question2: 数字乡村计划的主要目的是什么?(What is the main purpose of the Digital Rural Plan?)

Question Classification: normal

Question1: 大数据有哪些主要特征？请分析说明这些特征的含义。(What are the main characteristics of Big Data? Please analyze and explain the meanings of these characteristics.)

Question2: 社会信息道德包括哪三个层次?简要描述每个层次的主要内容。(What are the three levels of social information ethics? Briefly describe the main content of each level.)

Question Classification: difficult

Question1: 设计一个Python函数,接收梯形的上底长度、下底长度和高作为参数,返回梯形的面积。\n此外,在主程序中调用该函数,计算一个上底长为5cm、下底长为7cm、高为9cm的梯形的面积。\n请详细解释函数的作用,并展示完整的代码。(Design a Python function that receives the lengths of the top base, bottom base, and height of a trapezoid as parameters and returns the area of the trapezoid. Additionally, in the main program, call this function to calculate the area of a trapezoid with a top base length of 5 cm, a bottom base length of 7 cm, and a height of 9 cm. Please explain the function's purpose in detail and display the complete code.)

Question2: 在数据分析中，如何有效地结合数据可视化和公式计算来揭示数据的内在关系并支持决策制定？请详细说明如何利用电子表格中的公式和函数进行复杂的数据分析。(In data analysis, how can data visualization and formula calculations be effectively combined to reveal intrinsic relationships in the data and support decision-making? Please explain in detail how to use formulas and functions in spreadsheets for complex data analysis.)

Figure 14: Examples of questions of varying difficulty from the ID evaluation dataset, and two examples are provided for each category of questions.

Question Score: 10
Question: 你能识别出数据中的任何异常值吗？这些异常值可能的原因是什么？(Can you identify any outliers in the data? What might be the possible causes of these outliers?)

Question Score: 30
Question: 你能解释一下堆栈如何跟踪多个递归调用吗？(Can you explain how the stack tracks multiple recursive calls?)

Question Score: 55
Question: 有哪些创新方法可以利用并行计算和GPU加速来优化算法的性能？(What innovative methods can be used to optimize algorithm performance using parallel computing and GPU acceleration?)

Question Score: 70
Question: 栈如何帮助管理深度学习模型中的层和计算？(How does the stack help manage layers and computations in deep learning models?)

Question Score: 90
Question: 你将使用哪种类型的问题，例如选择题、开放式问题或评级量表问题？(What type of questions will you use, such as multiple-choice, open-ended, or rating scale questions?)

Question Score: 115
Question: 二叉搜索树在与搜索和排序相关的算法中扮演了什么角色，它们如何结合以优化性能？(What role do binary search trees play in search and sort algorithms, and how do they combine to optimize performance?)

Figure 15: Examples of questions from the OOD evaluation dataset categorized by Bloom's taxonomy scoring.

Book Segment: 数字化是将信息转变为可度量的数据，再将这些数据转换为计算机可以处理和存储的二进制编码。随着大数据、人工智能和物联网技术的发展，不仅事物、环境可以被数字化，而且人类的语言、行为、活动、思维都可以被数字化。例如，今天你坐了哪趟公交车、中午点了什么外卖、访问了哪些网页、在网上买了什么东西、走路的步数以及跑步时候的心率，这些行为都可以被数字化。数字化技术改变着人们的生活、工作和学习，推动着社会的发展和变革。(Digitalization involves converting information into measurable data, which is then transformed into binary code that computers can process and store. With the development of big data, artificial intelligence, and Internet of Things technologies, not only can objects and environments be digitalized, but human language, behavior, activities, and thoughts can also be digitized. For instance, today's activities such as which bus you took, what food you ordered for lunch, which websites you visited, what you purchased online, your walking steps, and your heart rate while running can all be digitalized. Digitalization technology is changing people's lives, work, and learning, driving social development and transformation.)
Summary:数字化将信息转为计算机处理的数据和编码，随着技术发展，生活中的事物、语言等被记录，改变人们的方式，推动社会发展。(Digitalization transforms information into data and codes that computers can process. As technology advances, various aspects of life, including objects and language, are being recorded. This shift changes how people live and drives societal development.)
Keywords:数字化;数据;信息(digitalization; data; information)

Figure 16: An example of generating summaries and keywords based on a book.

Question: 数据如何与人辩论？(How does data argue with people?)

baichuan7b-chat_book2qa_sft:数据不能直接与人辩论，因为数据本身没有意识和情感。但是，我们可以通过分析数据和生成数据来支持或反驳某个观点。例如，如果数据表明某种药物对治疗某种疾病有效，那么这些数据可以支持该药物的疗效；反之，如果数据表明该药物无效，则可能反驳该药物的有效性。此外，数据分析和可视化可以帮助人们更好地理解数据，从而做出更明智的决策。(Data cannot directly debate with people because data itself lacks consciousness and emotions. However, we can analyze and generate data to support or refute a particular viewpoint. For instance, if data indicates that a certain drug is effective in treating a disease, this data can support the drug's efficacy; conversely, if data shows the drug is ineffective, it can refute the drug's efficacy. Additionally, data analysis and visualization can help people better understand the data, leading to more informed decisions.)

baichuan7b-chat_book:作为一个大语言模型，我不能直接参与辩论。(As a large language model, I cannot directly participate in debates.)

baichuan7b-chat:作为一个大语言模型，我不能直接参与辩论。(As a large language model, I cannot directly participate in debates.)

Question: 数据如何帮助测量？(How can data assist in measurement?)

baichuan7b-chat_book2qa_sft:数据可以帮助测量，因为它们可以记录和量化各种现象。例如，在物理学中，数据可以用来测量物体的质量、速度、加速度等物理量；在生物学中，数据可以用来测量生物体的生长速率、基因表达水平等生理指标；在社会科学中，数据可以用来测量社会经济活动中的消费水平、人口密度等社会现象。\n\n此外，数据还可以用来进行数据分析，通过分析数据来发现规律、预测未来趋势以及优化决策过程。例如，在金融领域，数据可以用来预测股票价格走势；在医学领域，数据可以用来研究疾病的发展趋势和治疗效果；在交通领域，数据可以用来优化交通路线规划。(Data can aid in measurement by recording and quantifying various phenomena. For instance, in physics, data can be used to measure physical quantities such as an object's mass, velocity, and acceleration. In biology, data can measure physiological indicators like the growth rate of organisms and gene expression levels. In social sciences, data can measure social phenomena such as consumption levels and population density in socio-economic activities.\n\nAdditionally, data can be used for data analysis to discover patterns, predict future trends, and optimize decision-making processes. For example, in finance, data can predict stock price movements. In medicine, data can study the progression of diseases and the effectiveness of treatments. In transportation, data can optimize route planning.)

baichuan7b-chat_book:数据可以帮助测量各种事物，例如物体的位置、速度、加速度等。它还可以帮助我们了解事物的变化情况，以及它们之间的相互关系。此外，数据还可以帮助我们预测未来的趋势和模式，从而更好地理解世界。(Data can measure an object's position, velocity, and acceleration, understand changes and relationships, and predict future trends and patterns, enhancing our world understanding.)

baichuan7b-chat:数据可以帮助测量各种事物，例如物体的位置、速度、加速度等。它还可以帮助我们了解事物的变化情况，以及它们之间的关系。通过收集和分析数据，我们可以更好地理解世界，并做出更明智的决策。(Data can help measure various things, such as an object's position, velocity, and acceleration. It can also help us understand changes and relationships among things. By collecting and analyzing data, we can better understand the world and make more informed decisions.)

Figure 17: In the performance comparison on the OOD evaluation set, it is evident that the output of the baichuan7b-chat_book2qa_sft model is more detailed and accurate, significantly reducing the hallucination phenomena observed in the original model.