AI-empowered Intelligent Education: Question Generation based on LLMs

Kehan Zheng
Tsinghua University

Yida Lu
Tsinghua University

Wenjing Wu
Tsinghua University

Abstract

College students often struggle with grasping complex materials in courses like physics and engineering due to limited access to personalized practice. Current AI-based question generation systems cover shallow knowledge and superficial formats, limiting their effectiveness as learning tools. In this work, we propose an AI-powered teaching assistant based on ChatGLM, which leverage Retrieval-Augmented Generation (RAG) and Supervised Fine-Tuning (SFT) to automatically generate specialized exercise questions in University Physics and Chemical Engineering Thermodynamics, fulfilling personalized requirements and effectively assisting college students with their study. Utilizing a multi-dimensional evaluation framework, our results show that multi-level RAG achieves the best performance, significantly improving question relevance and quality over the baseline. SFT with reflection also enhances question quality but remains inferior to RAG. These findings demonstrate the practical value of our approach while highlighting the need for improved reasoning capabilities to generate more complex and challenging questions.

1 Introduction

College students often face challenges in getting enough personalized practice, which hinders their ability to thoroughly grasp complex materials and improve their academic performance. This issue is particularly pronounced in courses like physics and engineering, where students are required to develop a deep conceptual understanding alongside strong problem-solving skills.

As education technology rapidly advances, there is a trend to utilize AI to provide tailored learning experiences at scale, offering solutions to long-standing issues in education [5, 27]. However, existing methods struggle with covering logical knowledge implied in higher education, and fail to specialize in a certain subject, which results in insufficient capability as an AI educational assistant. To address these limitations, we propose an AI-powered teaching assistant based on ChatGLM [6], which offers personalized practice and learning resources for students. Our framework utilize RAG and SFT to generate customized questions and exercises aligned with specific topics or areas where a student may need additional practice, adapting to the individual needs of each user and thus creating a dynamic and interactive learning experience. Our framework is adopted as a subproject of the AI Teaching Assistant project at Tsinghua University.

In this work, we primarily focus on university-level courses such as University Physics and Chemical Engineering Thermodynamics, which require both theoretical knowledge and the ability to apply concepts in problem-solving scenarios. We collect exercises from the two courses with a cleaning process, and use GLM-4 to extract specific question generation requirements for each exercise problem, facilitating generation of question-answer pairs with unique requirements. Based on the refined dataset, we automatically generate exercise questions relevant to the given requirements with RAG and SFT, ensuring the high quality of questions and alignment with specific criteria.

Preprint. Under review.

To assess the effectiveness of our framework, we leverage a multi-dimensional evaluation framework derived from AlignBench [16, 25]. The experimental results demonstrate that our proposed RAG-based models achieve significant improvements over the baseline in all evaluation metrics, particularly when utilizing a multi-level retrieval mechanism. This highlights the importance of effectively retrieving relevant examples to guide question generation. In addition, the SFT models incorporating a reflection mechanism outperform the baseline as well, although they still fall short of the RAG performance. However, the relatively limited reasoning capability of the fine-tuned models remains a challenge in generating complex and cognitively demanding questions.

From a broader perspective, our findings underscore the practical value of integrating RAG and reflection-enhanced SFT techniques into AI-powered educational tools. While there is still room for improvement, particularly in generating more conceptually challenging questions, our work provides a solid foundation for future research. Directions such as integrating explicit reasoning chains, inspired by emerging models like the O1 series, may further unlock the potential of AI for intelligent question generation in education.

2 Related Work

LLMs for Text Generation Text Generation aims at producing coherent natural language responses to human input and is a fundamental task for language models [14]. With the advancement of machine learning techniques, large language models (LLMs) have achieved remarkable performance across multiple text generation tasks [3, 24, 4]. LLMs, such as GPT-4 [1], typically adopt a decoder-only architecture and employ next token prediction objective, which is well-suited for text generation. In consequence, these models maintain an outstanding capability on generation tasks. In contrast, GLM [6] unifies different pre-training architectures with autoregressive blank filling, demonstrating comparable performance to decoder-only models in text generation. To facilitate the application and development of LLMs, many open-sourced LLMs have emerged [19, 11] and are widely used for academic or commercial purpose. Our work leverages LLMs' impressive text generation ability and utilizes open-sourced LLMs as AI teaching assistants.

Methods for Specific Task Generation As LLMs are trained on general objectives, it is essential to align LLMs with specific downstream tasks. A common method involves prompting LLMs for task completion [8, 15]. However, improper prompts may result in suboptimal performance [12]. To better adapt LLMs to the downstream tasks, supervised fine-tuning (SFT) is often employed [18, 21, 22]. SFT leverages labeled data from a given task and trains an LLM with Cross Entropy loss. While SFT is effective in enhancing LLMs' task-specific performance, it demands a significant amount of high-quality data, which can be difficult to obtain in certain fields. Another approach is Retrieval-Augmented Generation (RAG), which assists LLMs with retrieved passages [9, 13, 7, 2]. RAG retrieves relevant documents according to the input and provides the results to LLMs, thereby facilitating more accurate responses. RAG improves the reliability of the generated results and reduces training costs. In this work, we primarily utilize RAG and SFT to align LLMs to teaching assistant tasks.

Applications of Question Generation Question generation has emerged as a critical tool in various real-world applications, significantly advancing industries such as education, customer engagement, and data collection. Tools like Kuangyou AI and Doubao question generation specialize in generating high-quality, subject-specific questions for personalized learning, automating tedious manual work for educators, and enabling adaptive learning systems. Platforms like Jinshuju AI question generation leverage question generation to streamline survey creation and data collection, while GPT-based applications such as OpenAI's advancements in natural language processing [3] expand the scope of question generation to include corporate training, customer support, and even gaming. Recent research further demonstrates the effectiveness of question generation in enhancing personalized education [27] and automating question creation from textual resources [5]. These innovations highlight the transformative potential of question generation in automating content creation, enhancing user engagement, and improving workflow efficiency, making it an essential component of intelligent systems in education, business, and research contexts.

However, current AI-based question generation systems have several limitations that hinder their broader application in education. One significant drawback is that these systems often cover only

shallow knowledge, focusing mainly on simpler topics and rarely addressing more complex, logical concepts necessary for higher education. Additionally, many existing question-generation tools remain at a superficial level, generating basic fill-in-the-blank or multiple-choice questions based on provided texts, without delving into subject-specific intricacies or being able to make inferences from the provided material to create questions on related topics. Furthermore, these systems are generally designed for business contexts, such as knowledge assessments or qualification testing, rather than for educational use. This limits their effectiveness as learning tools, as they are not tailored to enhance student comprehension or critical thinking skills.

3 Method

3.1 Definition

Question generation is a text generation task where, given specific requirements such as knowledge points, testing formats, and question types, a model is designed to produce questions that meet these criteria and can be answered effectively. However, current models often face significant issues, such as missing variables and poor association with the intended knowledge points. This project aims to improve the model's question-setting performance through various optimization techniques, enhancing its practical applicability.

3.2 Data

This study is part of a subproject of the AI Teaching Assistant project at Tsinghua university. For our study, we select two primary courses from the project: **University Physics** and **Chemical Engineering Thermodynamics**. We are provided with a total of 2124 exercises from the University Physics course and 2942 exercises from the Chemical Engineering Thermodynamics course. These exercises are initially extracted through OCR (Optical Character Recognition) from the course textbooks.

To improve the quality of the data, we conduct a thorough cleaning process. This involves removing exercises with incomplete problem statements, ensuring that only fully formed questions remain for further analysis. Additionally, as our primary focus is on pure-text type questions, we filter out exercises that require images or other non-textual content. After the cleaning process, we are left with 1937 exercises from the University Physics course and 2344 exercises from the Chemical Engineering Thermodynamics course.

Once we have the clean dataset of exercises, we use the GLM-4 model to extract the specific question generation requirements for each problem. This step allows us to generate question-answer pairs consisting of the question requirements (i.e., the specific task the model needs to fulfill) and the corresponding questions themselves.

This refined dataset serves as the foundation for our further research on automated question generation. The data follows the format described below:

- **User Question** (`question`): This field contains a string-formatted query provided by the user to generate a problem. For example: "Please generate a problem on the quantum mechanics infinite square well in one dimension."
- **Reference Answer** (`reference`): This field contains a string-formatted correct problem extracted from the *University Physics* dataset. For example: "Problem: Find the series representation of $\psi(x, t > 0) \ldots$"

To evaluate the quality of the generated questions, we randomly select 100 questions from each subject as a test set. Due to intellectual property concerns, all questions will not be disclosed.

3.3 Metrics

AlignBench enables the automated evaluation of large language models (LLMs) through customizable evaluation strategies tailored to real-world scenarios[16, 25]. By integrating comprehensive multi-dimensional benchmarks with a robust data management process, AlignBench employs a rule-calibrated multi-dimensional *LM-as-Judge* approach. This is combined with chain-of-

thought prompting to generate evaluation rationales and final scores, ensuring high reliability and interpretability[26, 20].

In our question-generation scenario, the evaluation task focuses on assessing the quality of problems generated by the system in response to user requirements. The evaluation includes the following four dimensions:

- **Relevance to Knowledge Points**: Measures whether the generated problem aligns with the knowledge points specified in the user query.
- **Problem Correctness**: Evaluates whether the generated problem is correct and solvable, encompassing three aspects: Correctness of formulas, Solvability of the question format, and Accuracy of hints and solutions.
- **Problem Completeness**: Assesses whether the problem is independent and complete, such that no additional conditions are required during the solving process. The problem should include necessary parameters, assumptions, and be more specific than the user-provided prompt.
- **Problem Difficulty**: Evaluates whether the problem difficulty meets the standards of university-level coursework. This includes: Covering multiple knowledge points and common pitfalls, and Involving complex logic and calculations in the solution process.

This work adapts AlignBench's multi-dimensional evaluation framework, initially designed for assessing the overall performance or specific capabilities of large language models, to the task of evaluating question quality. At its core is the instruction design for the rule-calibrated multi-dimensional *LM-as-Judge*, which includes a description of task instructions, the definition of evaluation dimensions, the evaluation process, and the scoring criteria.

In the question quality assessment scenario we have designed, the task instruction requires the judge to evaluate the quality of the questions generated by our question-generation assistant based on user needs. The evaluation dimensions include four aspects: relevance to the given knowledge points, problem correctness, problem completeness, and problem difficulty. The evaluation process involves comparing the assistant's output with the reference answer, evaluating the assistant's response across these dimensions, and scoring the response according to these dimensions. The scoring system is divided into five levels, ranging from 1 to 10 points.

The evaluation process requires concatenating the 'question', 'reference', and 'answer' fields from the dataset with the judge's prompt instructions, and then inputting this combined information into the large language model judge. The judge then evaluates and assigns a score to the response.

In this study, we select the GLM-4 model as our base model. Given that this model has demonstrated strong text generation capabilities, it serves as an ideal starting point for our research. Our research also builds upon GLM-based models, leveraging their strengths to further refine and improve automated question generation methods.

3.4 Framework Design

3.4.1 Retrieval Augmented Generation

To enhance the precision and pertinence of generated question, we propose a RAG system with a multi-level retrieval mechanism, ensuring a wide coverage of knowledge point and topic requirement. The framework of our system is presented in Figure 1. The knowledge points and problem texts are first transformed into high-dimension vectors with an M3E embedding model [23] to extract semantic information. Based on this process, we construct corresponding vector databases utilizing ChromaDB. For problems with images, we upload the images to a specially designed image database to provide a more comprehensive support for question generation.

Given the user input query comprising a series of knowledge queries and a requirement query, we employ the first-level retrieval and calculating the distances between these queries and the stored knowledge points using L2 norm, selecting the knowledge point with the shortest distance to each of them. From this procedure, a list of knowledge points are extracted from the database, and we further retrieve 20 problem texts with the minimal L2-norm distance to the requirement query, provide that at least one knowledge point of the problem text falls into the extracted knowledge points list. These

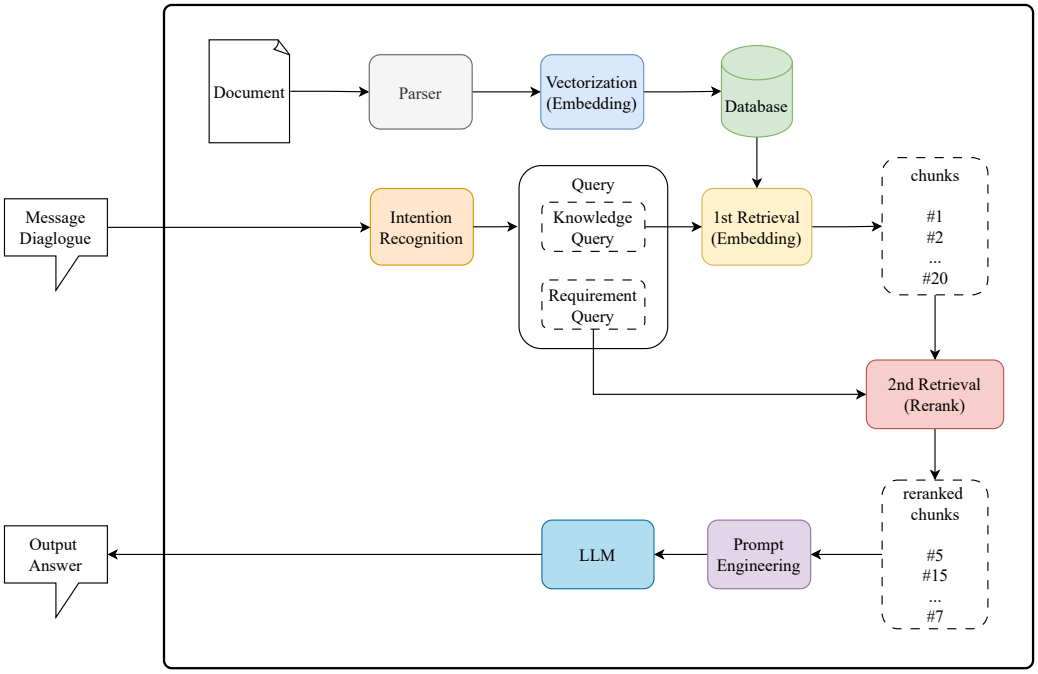

Figure 1: The framework of RAG system.

problem texts are regarded as the most relevant documents for the user input, and are sequentially taken as input for the second-level retrieval.

The second-level retrieval starts by re-ranking the retrieved documents according to the user input query with a BCE Rerank model [17]. Documents with high ranks are then leveraged by LLMs to generate questions that fulfill the requirements of the query.

Through multi-level retrieval, the RAG system effectively incorporate knowledge from different subjects with LLMs' outstanding capability, therefore improving the relevance, precision as well as difficulty of the questions. We also implement a single-level retrieval mechanism that contains only M3E to compare its performance with the multi-level type.

3.4.2 Supervised Fine-tuning

In the Supervised Fine-tuning section, we use the base model GLM-4-9B-Chat and fine-tune it using Llama-Factory with LoRA (Low-Rank Adaptation) [10]. For this, we set the input to be the question generation requirements and the output to be the generated questions. We perform fine-tuning on the model for two different subjects.

To further enhance the fine-tuning process, we explore several improvements. The first improvement involves incorporating scoring standards and desired output scores into the input data. Specifically, we design a scoring mechanism where each existing question in the dataset is evaluated using GLM-4 to assign a quality score based on criteria such as completeness, relevance, and clarity. This generated score is then included in the input as part of the supervised fine-tuning (SFT) training process. By embedding these quality scores into the training data, we aim to help the model better understand what constitutes a high-quality question and prioritize these characteristics during question generation.

The second improvement introduces a reflection mechanism to further refine the generated questions. After completing the SFT training, the model is prompted to assess and revise its own generated questions. Specifically, after generating a question, we continue the interaction by asking the model to evaluate its output from multiple dimensions, such as the completeness of the question setup, relevance to the associated knowledge points, and overall quality. Based on this evaluation, the model is then encouraged to modify and improve the question. This iterative process allows the

model to reflect on its outputs and address potential flaws, resulting in more refined and contextually appropriate questions.

These approaches aim to refine the model's question generation abilities, focusing not only on generating relevant questions but also on enhancing the quality and alignment with specific criteria.

4 Results

Based on the given evaluation metrics, we conduct tests on a test set of 200 questions for six different approaches: Baseline, RAG (single-level), RAG (multi-level), SFT (only), SFT (with-score), and SFT (with-reflection). The Baseline and RAG models both use the GLM-4 base model, accessed through an API, while the SFT models are based on the fine-tuned GLM-4-9B model. The testing results are shown in Table 1 and Table 2.

Table 1: Evaluation Results for University Physics of different models

	Relevance	Correctness	Completeness	Difficulty	Overall Score
Baseline	7.85	5.93	6.65	6.49	6.21
RAG (single-level)	**7.98**	6.20	<u>6.69</u>	**6.60**	<u>6.31</u>
RAG (multi-level)	<u>7.92</u>	**6.33**	**6.72**	<u>6.58</u>	**6.34**
SFT (only)	7.60	5.91	6.55	6.30	6.10
SFT (with-score)	7.58	5.93	6.57	6.24	6.12
SFT (with-reflection)	7.88	<u>6.23</u>	6.68	6.50	6.27

Table 2: Evaluation Results for Chemical Engineering Thermodynamics of different models

	Relevance	Correctness	Completeness	Difficulty	Overall Score
Baseline	7.67	6.02	6.67	6.43	6.23
RAG (single-level)	<u>7.82</u>	<u>6.27</u>	<u>6.73</u>	<u>6.45</u>	<u>6.31</u>
RAG (multi-level)	**7.93**	**6.53**	**6.77**	**6.48**	**6.47**
SFT (only)	7.50	5.95	6.55	6.30	6.10
SFT (with-score)	7.55	5.92	6.50	6.35	6.12
SFT (with-reflection)	7.72	6.20	6.68	6.40	6.25

From the tables, it can be observed that the RAG approaches outperform the Baseline across all evaluation metrics, with the multi-level RAG showing further improvements over the single-level RAG. This indicates that effectively retrieving relevant examples or context aids the model in generating higher-quality questions. The multi-level retrieval mechanism refines this process, leading to improved relevance, correctness, and overall score, which demonstrates the benefits of leveraging more structured and multi-layered retrieval.

In contrast, the performance of the SFT (only) model is inferior to the Baseline. Upon inspecting the generated results, several issues are identified, such as missing variables, overly simplistic question setups, and a lack of contextual diversity. Even when incorporating score guidance in the SFT (with-score) model, only marginal improvements are observed compared to the SFT (only) model. This limited impact may be attributed to the way scoring information is utilized in the training data. Specifically, in the current dataset, each question generation requirement is paired with only a single score, which does not provide the model with a comparative understanding of good versus bad questions. Without exposure to contrasting examples, the model lacks the ability to discern how different quality levels influence the scoring criteria, thereby limiting its capacity to leverage this information effectively during fine-tuning.

For the SFT (with-reflection) model, where the model is prompted to identify and correct potential flaws in its generated questions, there is a noticeable improvement compared to the Baseline. However, it still underperforms relative to the RAG approaches. Analysis of the results reveals that while the reflection mechanism enables the model to detect and address certain problems in the generated questions, it often fails to identify issues even when explicitly prompted to reflect, resulting in limited corrections and improvements.

Overall, the RAG approaches demonstrate a clear advantage in improving question generation quality, particularly through the use of relevant retrieval mechanisms. The introduction of reflection in the

SFT models leads to some gains beyond the Baseline, but these remain insufficient to match the performance of the RAG approaches. This discrepancy may stem from the fact that the RAG methods utilize a base model that has undergone domain-specific fine-tuning, giving it a natural edge over the open-source GLM-4-9B-Chat model used in the SFT methods. However, since the GLM-4 model is not open-sourced, further fine-tuning or experimentation on it remains infeasible, which limits the ability to explore its full potential for question generation.

5 Discussion

In this work, we designed a question generation framework for two courses, namely University Physics and Chemical Engineering Thermodynamics. We constructed corresponding datasets and defined evaluation metrics to assess the quality of generated questions. The framework was built from two directions: Retrieval-Augmented Generation and Supervised Fine-Tuning . The evaluation results demonstrate that both multi-level RAG and SFT with a reflection mechanism outperform the baseline model. Among these, the multi-level RAG achieves the best results, indicating that our proposed methods hold strong practical value and can effectively address some of the existing issues in question generation.

However, from the evaluation results, it is clear that the current question generation outcomes still have significant room for improvement. Most of the generated questions are relatively straightforward and can be solved by applying basic formulas, lacking sufficient cognitive depth and complexity. This limitation highlights the challenge of generating questions that require higher-order thinking. To achieve this, models would need enhanced reasoning capabilities; if a model cannot solve difficult questions, it would be equally challenging to generate them effectively.

Due to time constraints in the course, we explored only two major directions: RAG and SFT. In fact, with the emergence of advanced models such as the O1 series, slow and deliberate reasoning enabled by long-chain thinking presents another promising avenue for question generation. Future research can focus on constructing explicit thought chains tailored for question generation, providing the model with improved reasoning paradigms to reference. This approach has the potential to maximize the model's capacity to generate higher-quality and more complex questions, unlocking its full potential in automated question design.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[2] Boros, E., et al., 2024. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In: Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP). pp. 75–82.

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

[4] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al., 2023. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research 24 (240), 1–113.

[5] Du, X., Shao, J., Cardie, C., 2017. Learning to ask: Neural question generation for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1342–1352.

[6] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J., 2022. Glm: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 320–335.

[7] Gao, S., Xiong, C., Gao, J., et al., 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

[8] Gao, T., Fisch, A., Chen, D., 2021. Making pre-trained language models better few-shot learners. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3816–3830.

[9] Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.-W., 2020. Realm: retrieval-augmented language model pre-training. In: Proceedings of the 37th International Conference on Machine Learning. pp. 3929–3938.

[10] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[11] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

[12] Jiang, Z., Xu, F. F., Araki, J., Neubig, G., 2020. How can we know what language models know. Transactions of the Association for Computational Linguistics 8, 423–438.

[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Advances in Neural Information Processing Systems. Vol. 33. pp. 9459–9474.

[14] Li, J., Tang, T., Zhao, W. X., Nie, J.-Y., Wen, J.-R., 2024. Pre-trained language models for text generation: A survey. ACM Computing Surveys 56 (9), 1–39.

[15] Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., Tang, J., 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68.

[16] Liu, X., Lei, X., Wang, S., Huang, Y., Feng, Z., Wen, B., Cheng, J., Ke, P., Xu, Y., Tam, W. L., Zhang, X., Sun, L., Wang, H., Zhang, J., Huang, M., Dong, Y., Tang, J., 2023. Alignbench: Benchmarking chinese alignment of large language models. arXiv preprint arXiv:2311.18743.

[17] NetEase Youdao, I., 2023. Bcembedding: Bilingual and crosslingual embedding for rag.

[18] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems 35, 27730–27744.

[19] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[20] Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., Sun, H., 2022. Towards understanding chain-of-thought prompting: An empirical study of what matters. arXiv preprint arXiv:2212.10001.

[21] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., Hajishirzi, H., 2023. Self-instruct: Aligning language models with self-generated instructions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13484–13508.

[22] Wang, Y., Li, W., Zhang, L., et al., 2024. Supervised fine-tuning achieves rapid task adaptation in large language models. arXiv preprint arXiv:2409.15820.

[23] Wang Yuxin, Sun Qingxuan, H. s., 2023. M3e: Moka massive mixed embedding model.

[24] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al., 2022. Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414.

[25] Zhang, W., Chen, L., Yan, M., et al., 2023. Alignmmbench: Evaluating chinese multimodal alignment in large vision-language models. arXiv preprint arXiv:2308.14353.

[26] Zhang, Z., Zhang, A., Li, M., Smola, A., 2022. Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493.

[27] Zhou, Q., Yang, N., Wei, F., Zhou, M., 2017. Neural question generation from text: A preliminary study. Natural Language Processing and Chinese Computing, 662–671.