Enhanced Fine-Tuning of Lightweight Domain-Specific Q&A model Based on Large Language Models

Shenglin Zhang [1], Pengtian Zhu [1], Minghua Ma [2], Jiagang Wang [3], **Yongqian Sun *** [1], Dongwen Li [1], Jingyu Wang [1], Qianying Guo [4], Xiaolei Hua [4], Lin Zhu [4], Dan Pei [3]

[1] Nankai University, [2] Microsoft,
[3] Tsinghua University, [4] China Mobile Research Institute

Outline

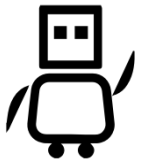



Background Design Evaluation Conclusion

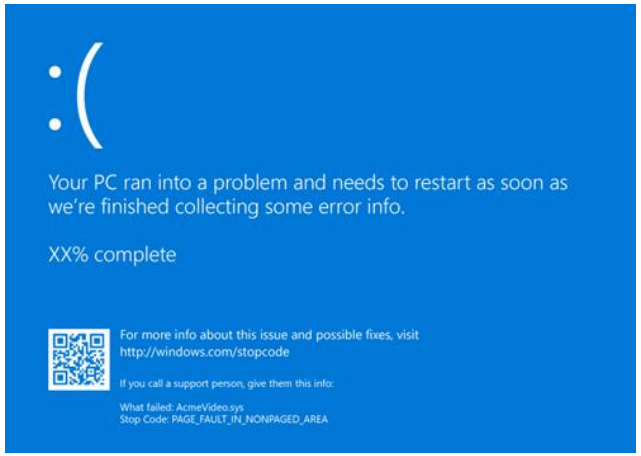
Domain-Specific LLM

LLMs can effectively answer questions in the public domain

My Windows has crashed with error code xxx, how do I fix it?

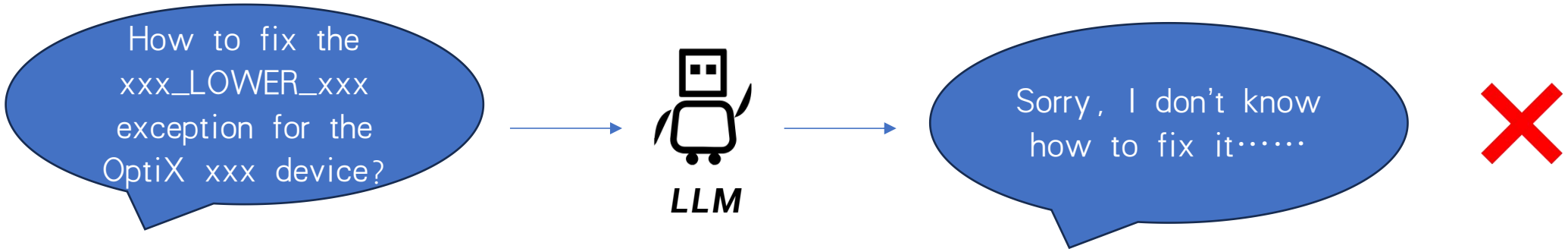


Step 1. Right-click the Start button and choose Settings.
Step 2. ……



Domain-Specific LLM

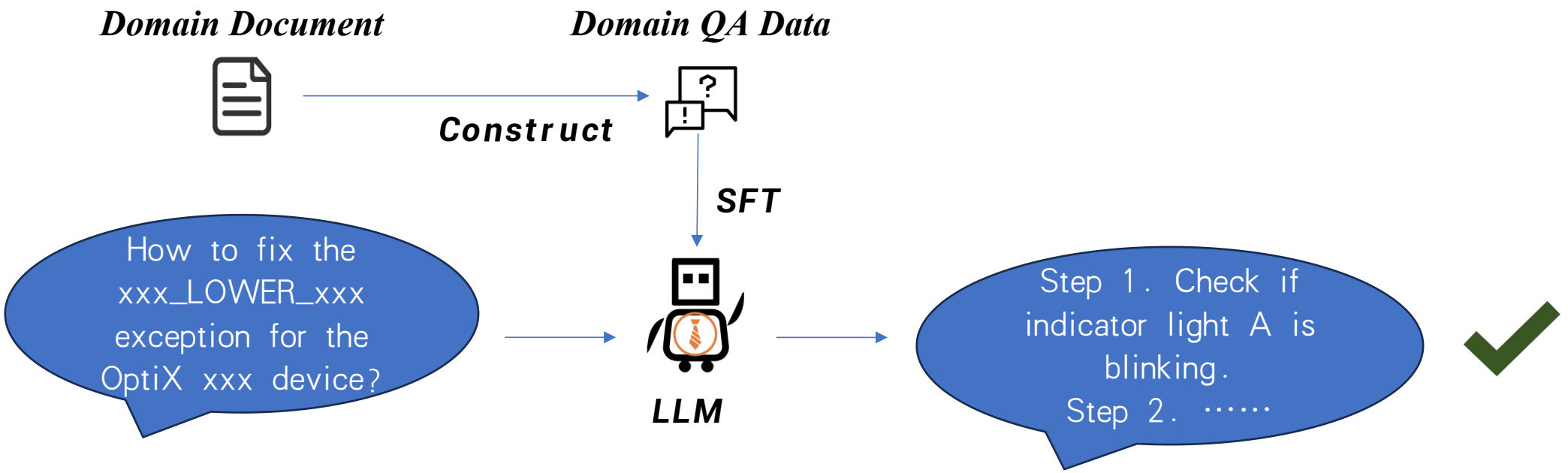
For specific domain questions, they often perform poorly



Domain-Specific Q&A LLM is needed.

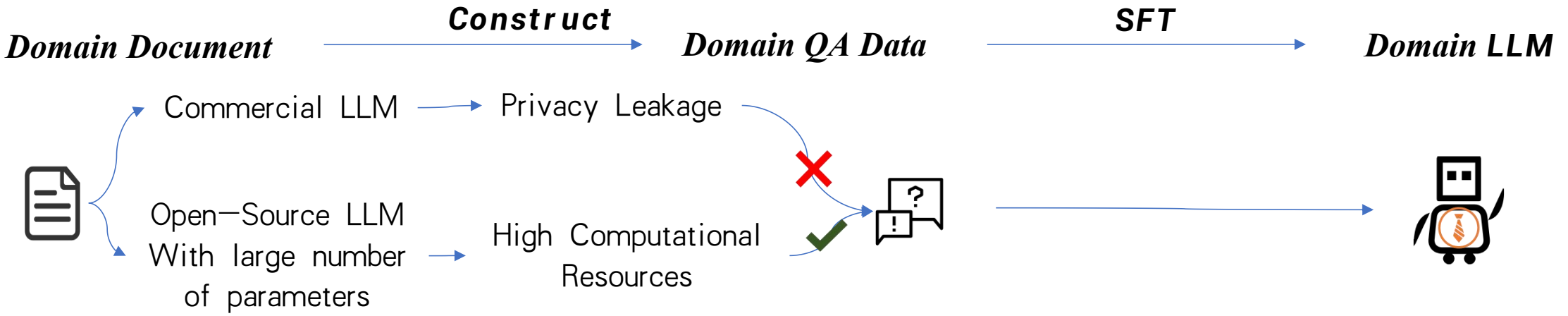
Domain-Specific LLM

We can achieve better results through SFT (Supervised Fine-Tuning).



High-Quality Instruction Data

The amount of document often greatly exceeds the quantity of instruction data



You need either **5 × V100** or **2 × A100** to run a 72B LLM, which is very expensive.

But can the same effect be achieved within **1 × V100** (using a 7B LLM)?

Challenges

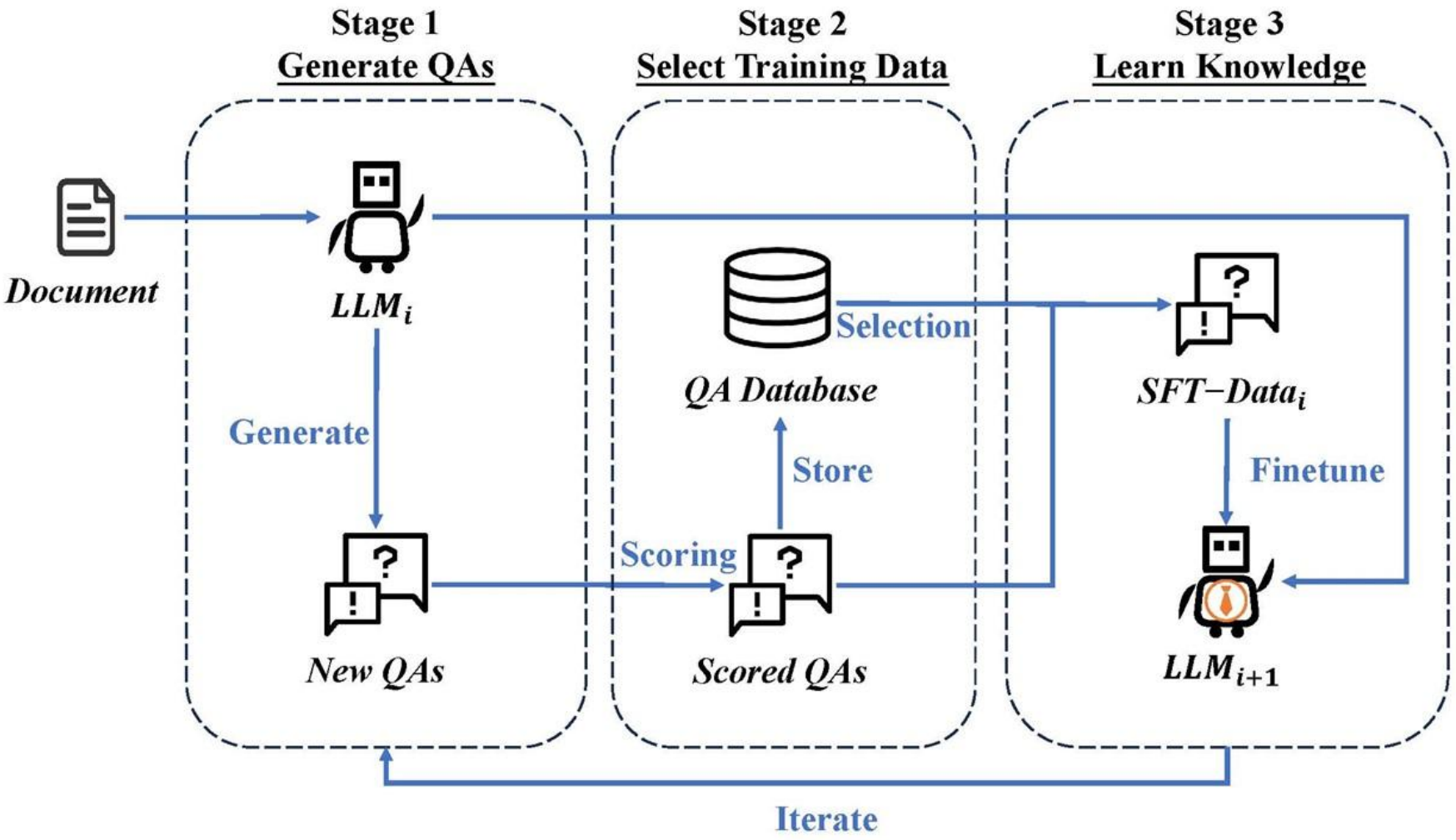
- Require high-quality instruction data from the specific domain documents
- The need to protect data privacy
- The high cost of resources required to construct instruction data

Outline

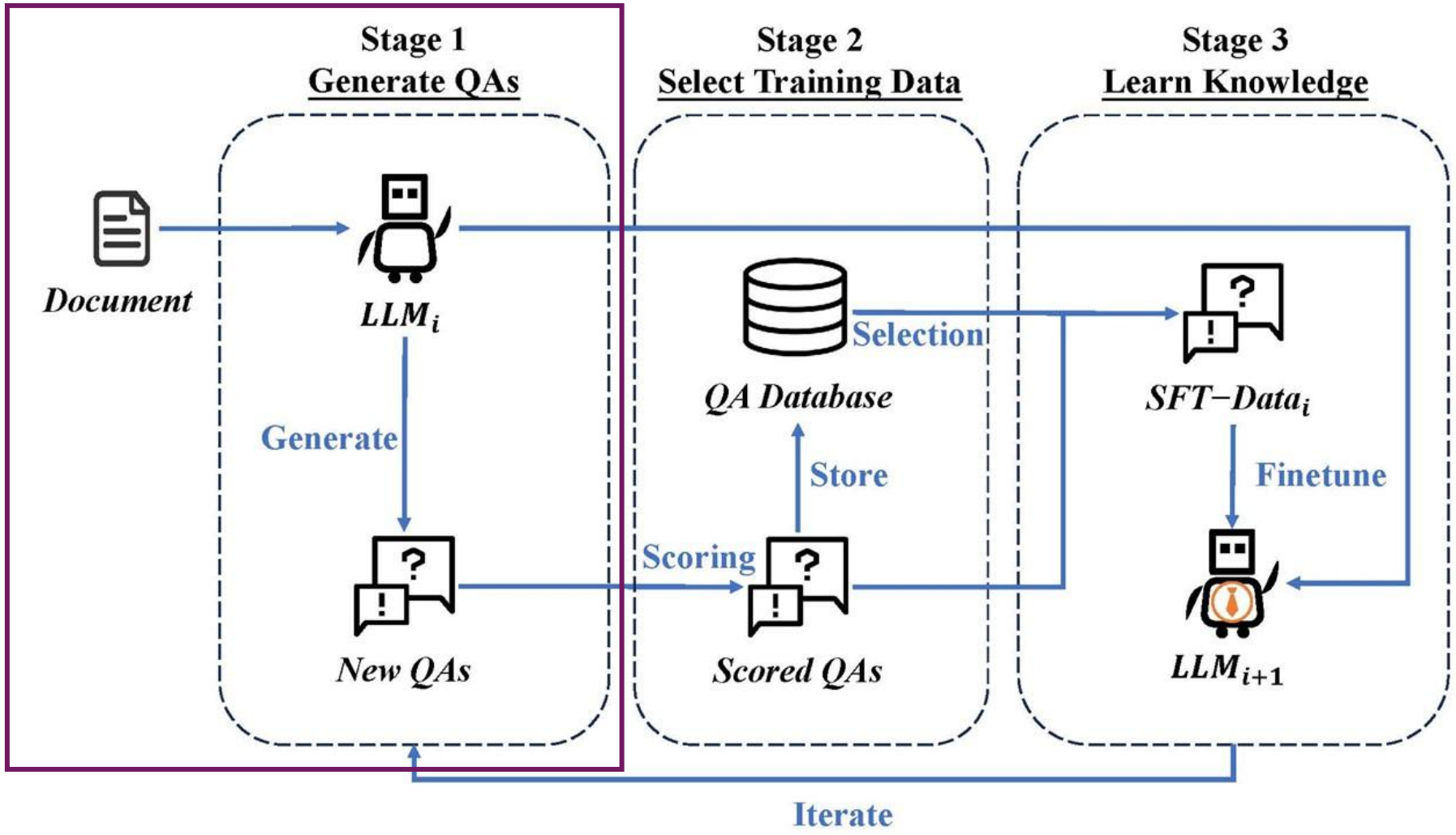



Background **Design** Evaluation Conclusion

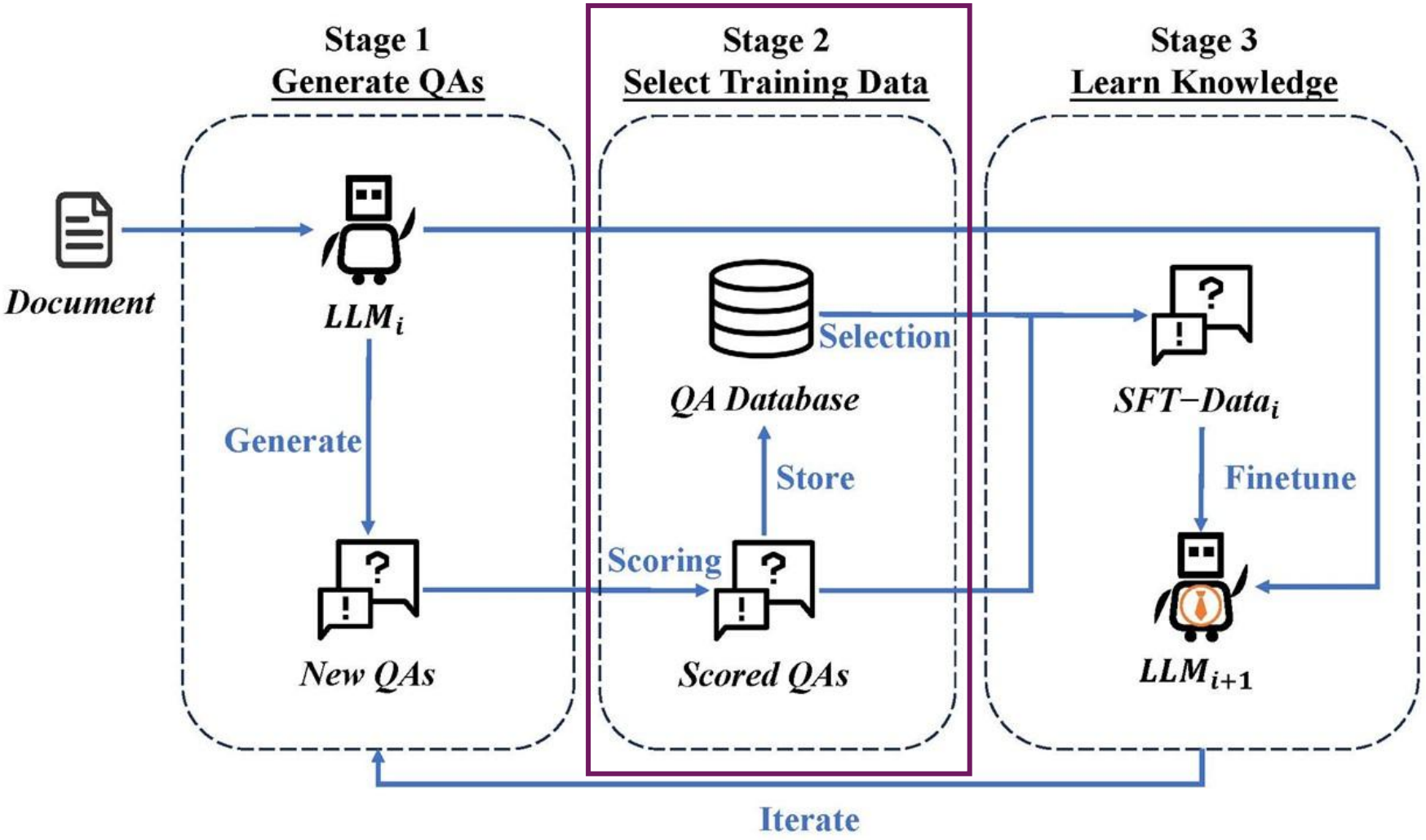
Self-Evolution Overview



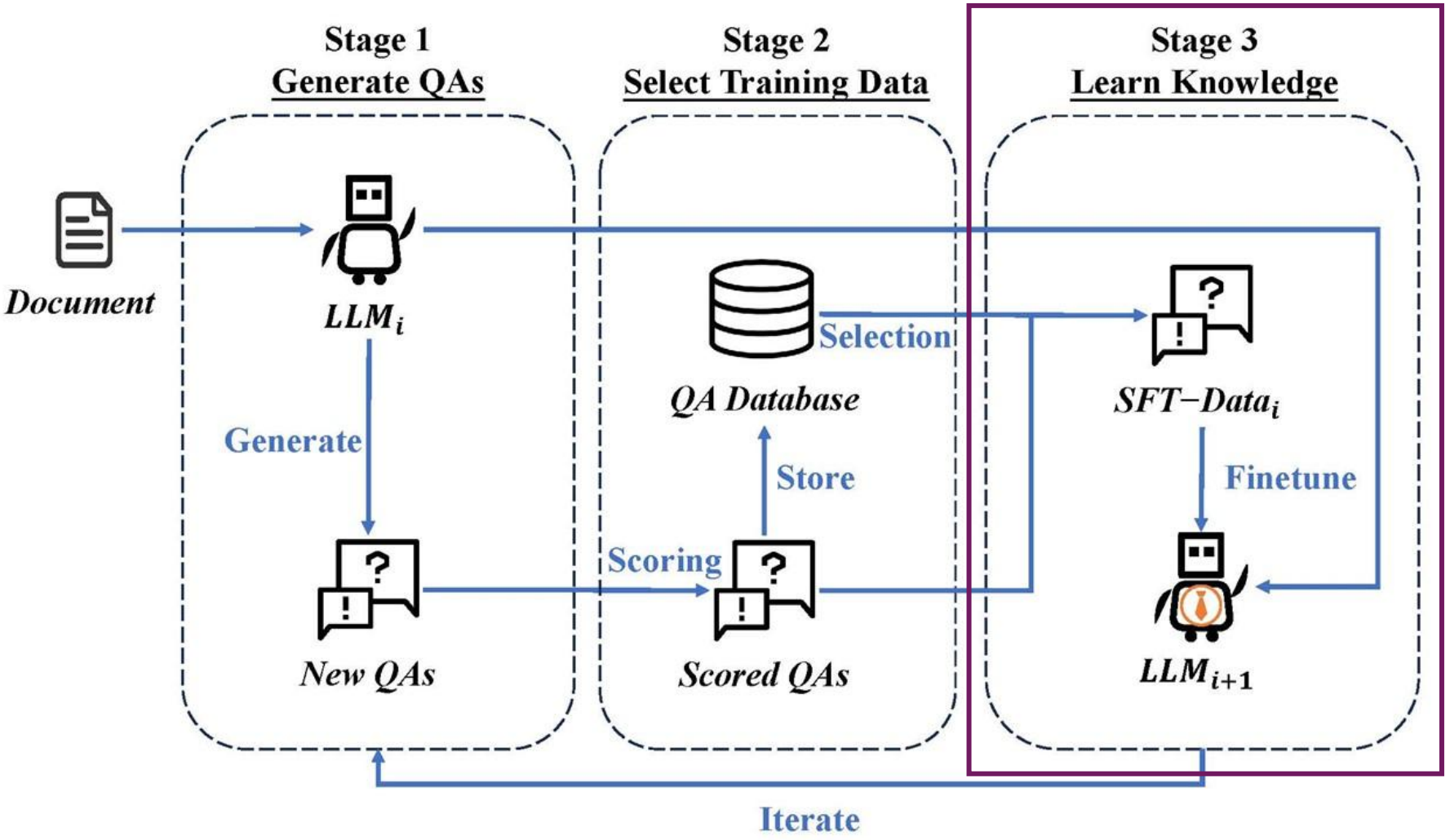
Self-Evoluation Overview



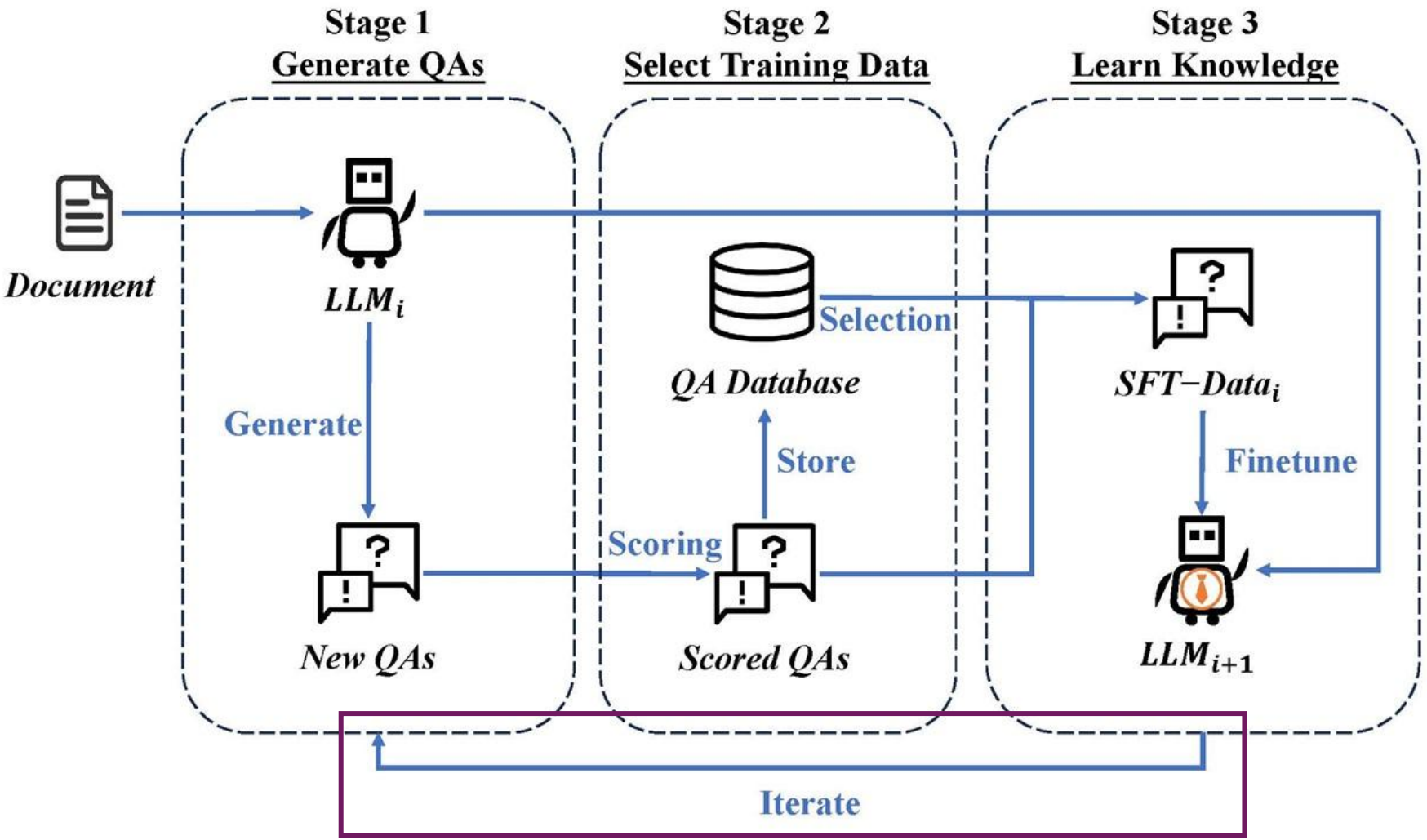
Self-Evolution Overview



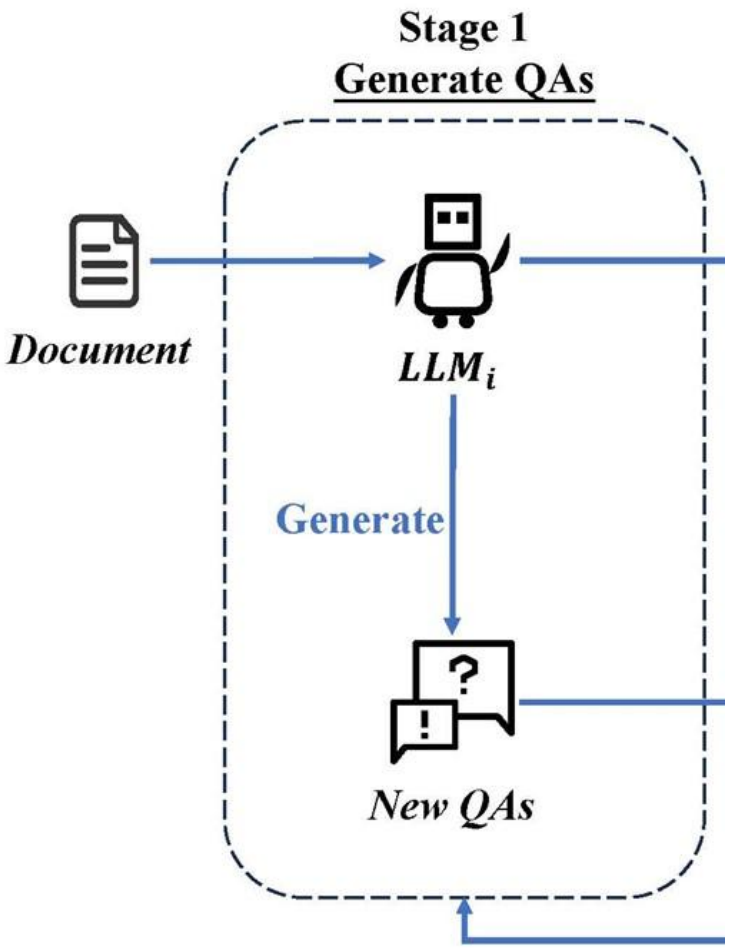
Self-Evolution Overview



Self-Evolution Overview

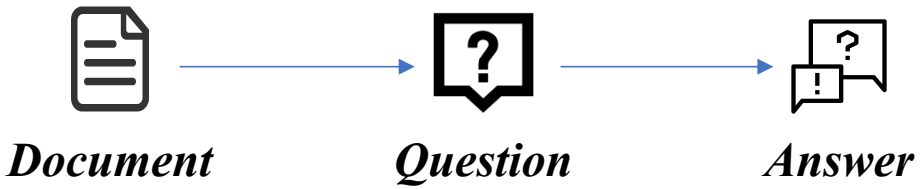


Generate QAs

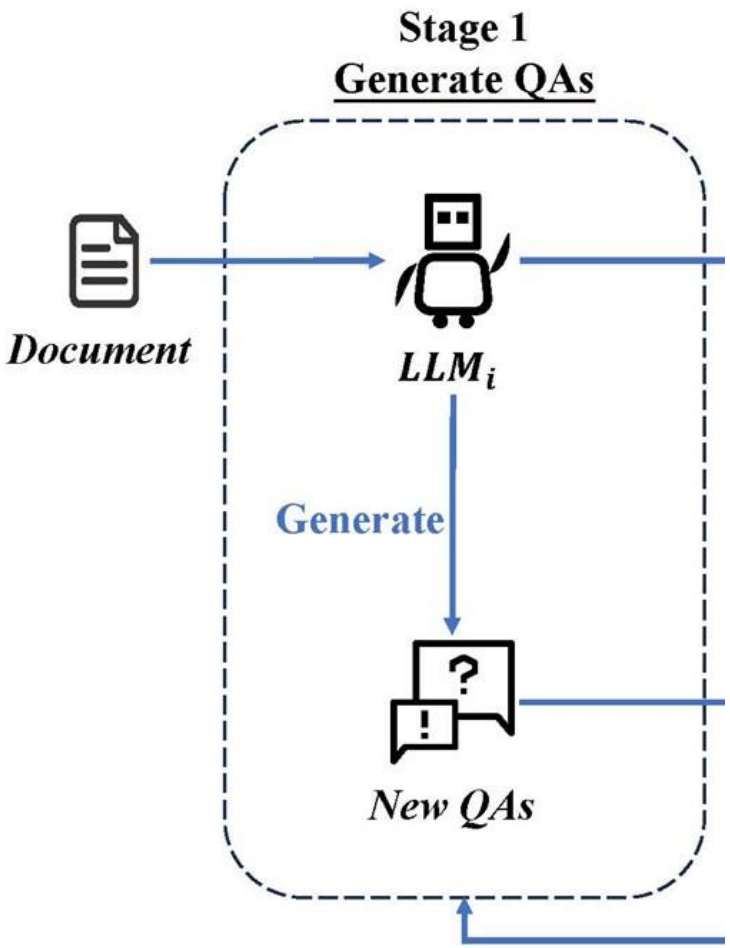


Two–Stage Inference

- Reference documents are often lengthy.
  There is a conflict in the intent of prompt design.
  The generation format is unstable.



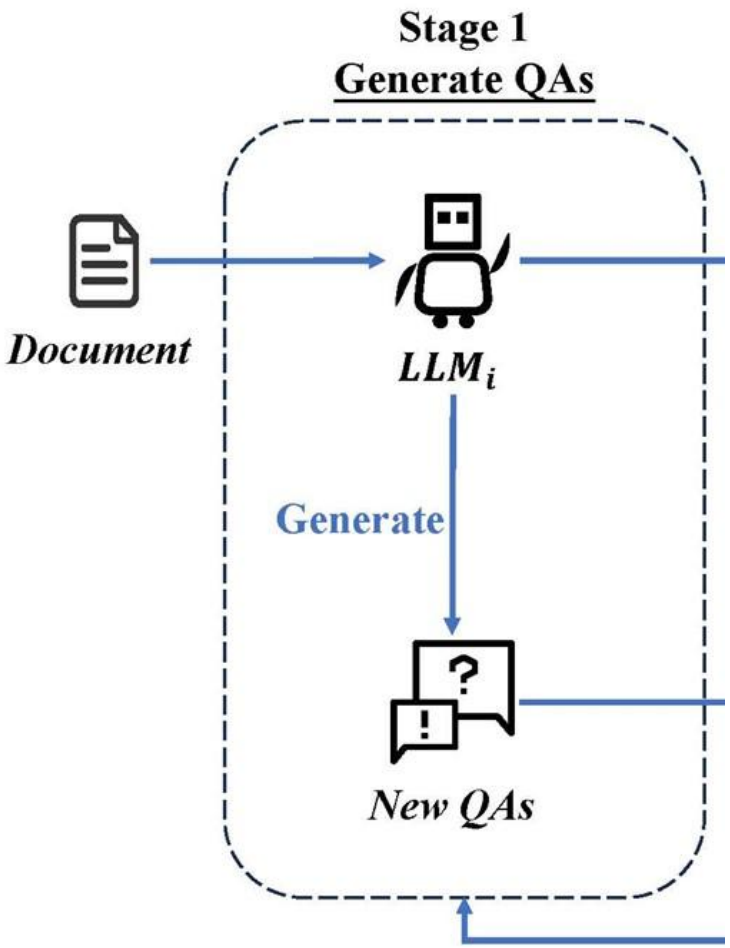
Generate QAs



Domain Knowledge:
Reference document: {Knowledge}
Role Description:
You are an expert in the operations domain.
Based on your comprehensive knowledge and the information provided above.....
Rules Description:
Note 1: The question should be as concise as possible.
Note 2: The question should not contain multiple sub-questions, only one question is permitted.
.....
Note 6: Do not output declarative sentences; it must be a question!
Please formulate a question now.
Question:

Question Generation Prompt

Generate QAs



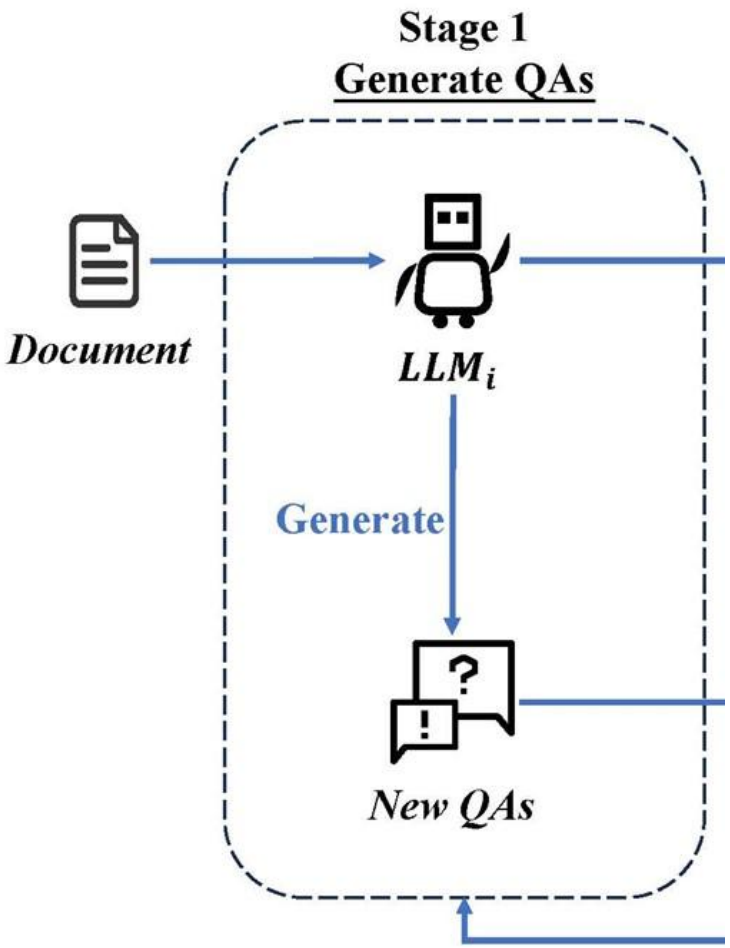
Domain Knowledge:
Reference document: {Knowledge}
Role Description:
You are an expert in the operations domain.
Based on your comprehensive knowledge and the information provided above.....
Rules Description:
Note 1: The question should be as concise as possible.
Note 2: The question should not contain multiple sub-questions, only one question is permitted.
.....
Note 6: Do not output declarative sentences; it must be a question!
Please formulate a question now.
Question:

Question Generation Prompt

What is a process, how to achieve process synchronization, and how to evaluate the efficiency of process management?

Undesirable Question

Generate QAs



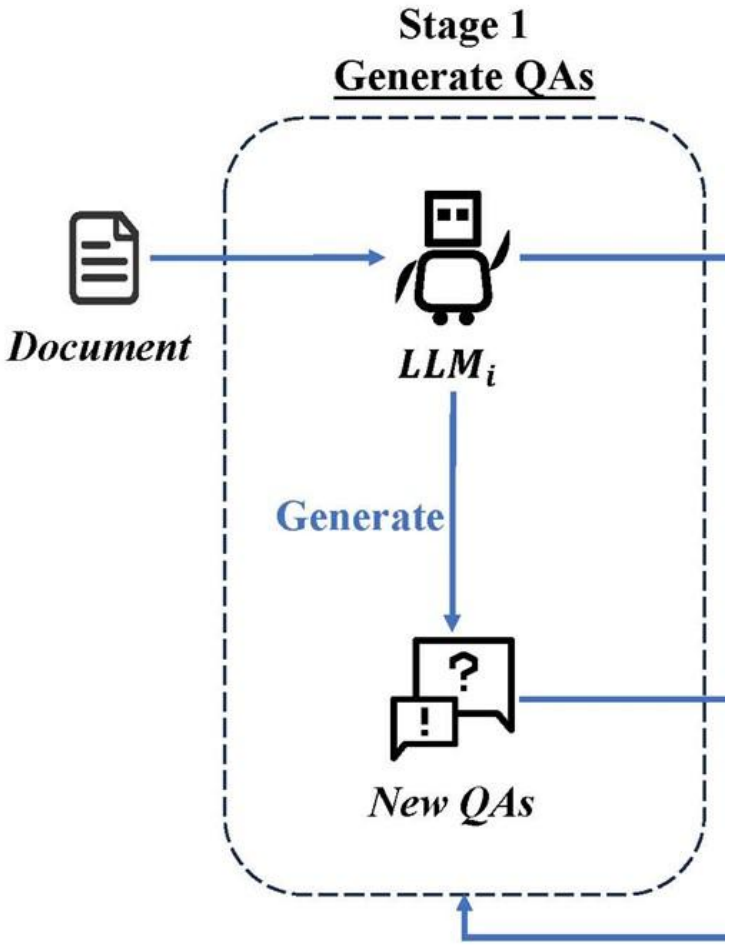
Domain Knowledge:
Reference document: {Knowledge}
Role Description:
You are an expert in the operations domain.
Based on your comprehensive knowledge and the information provided above.....
Rules Description:
Note 1: The question should be as concise as possible.
Note 2: The question should not contain multiple sub-questions, only one question is permitted.
.....
Note 6: Do not output declarative sentences; it must be a question!
Please formulate a question now.
Question:

Question Generation Prompt

What is a process? A process is the fundamental unit for resource allocation and scheduling in an operating system.

Undesirable Question

Generate QAs



Role Description:
You are an expert in the field of operations.....
You must generate responses based on the requirements.
Workflow Description:
1. Receive and parse the user's question.
2. Read and analyze the document provided by the user.
3. Provide a concise and comprehensive answer by combining your knowledge with the document content.
In Context Learning:
Examples:
Question: Which is the largest planet in the solar system?
Knowledge fragment: The solar system consists of eight planets, with Jupiter being the largest. Its mass is 2.5 times that of all other planets combined.
Answer: The largest planet in the solar system is Jupiter.
Warnings:
Your answer will be sent independently of the document after generation.....
Your response must ensure two points: conciseness and accuracy.
Domain Knowledge and Question:
Question: {Question}
Knowledge fragment: {Knowledge}

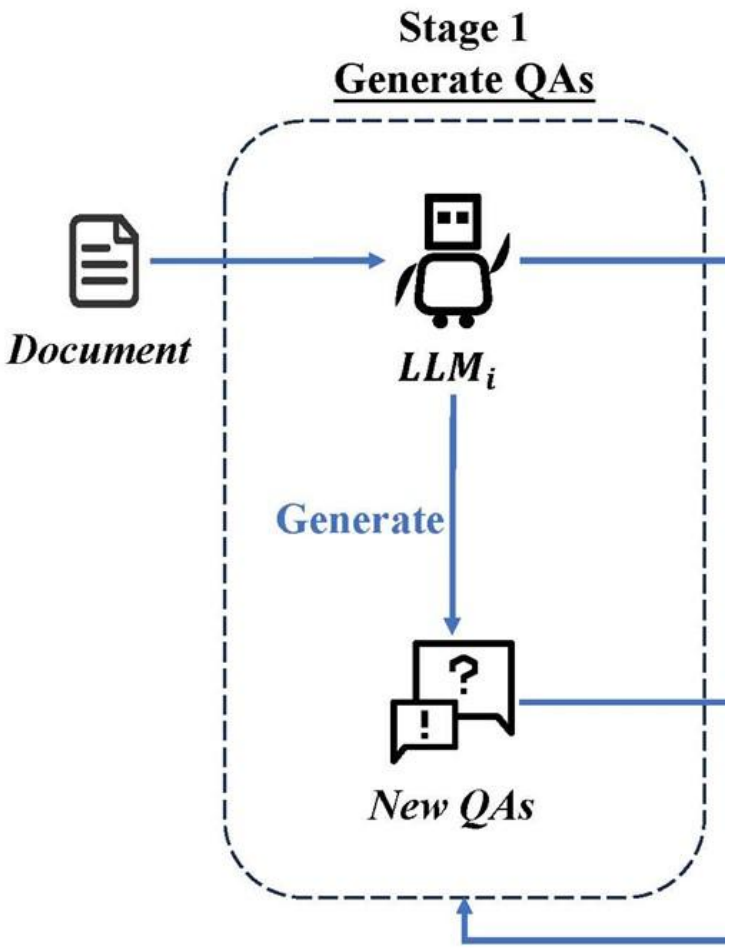
Answer Generation Prompt

Generate QAs

Q: The OptiX_xxx has reported an xxx_LOWER_xxx exception, and the indicator light is showing red. How should I fix this?

Answer A	Answer B ✓
According to the knowledge in the documentation, an engineer needs to replace the power supply.	When OptiX_xxx reports an xxx_LOWER_xxx exception, an engineer should check the color of the indicator light. 1. If the color is red, it indicates insufficient power supply, and the power supply needs to be replaced; 2. If the color is green, it means the current exception can be ignored; 3. If the color is yellow, it means the machine needs to be restarted promptly. 4. Since the indicator light is red, the engineer needs to replace the power supply with a new one as soon as possible.

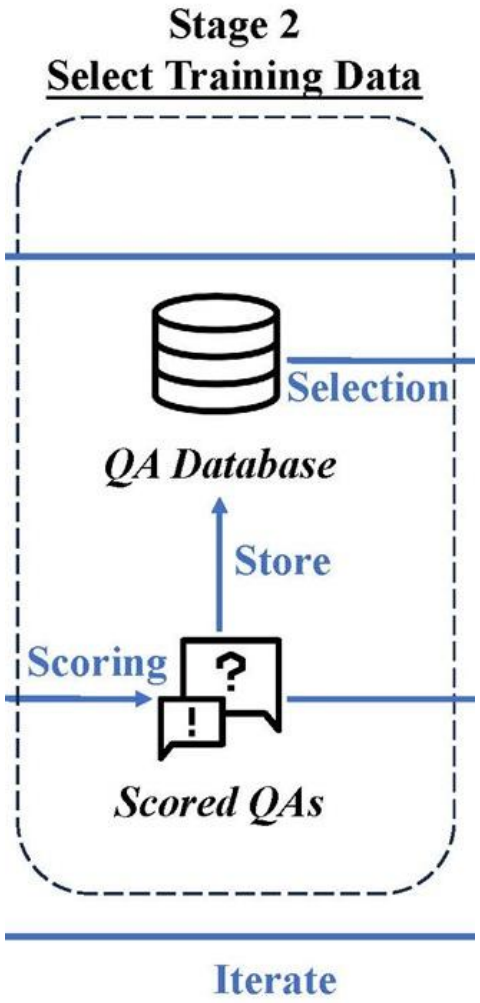
Generate QAs



Role Description:
You are an expert in the field of operations.....
You must generate responses based on the requirements.
Workflow Description:
1. Receive and parse the user's question.
2. Read and analyze the document provided by the user.
3. Provide a concise and comprehensive answer by combining your knowledge with the document content.
In Context Learning:
Examples:
Question: Which is the largest planet in the solar system?
Knowledge fragment: The solar system consists of eight planets, with Jupiter being the largest. Its mass is 2.5 times that of all other planets combined.
Answer: The largest planet in the solar system is Jupiter.
Warnings:
Your answer will be sent independently of the document after generation.....
Your response must ensure two points: conciseness and accuracy.
Domain Knowledge and Question:
Question: {Question}
Knowledge fragment: {Knowledge}

Answer Generation Prompt

Select Training Data



Instruction Following Difficulty[1] (IFD)

- To make full use of the previously generated QA data
- A higher IFD score indicates that the question is more difficult to answer

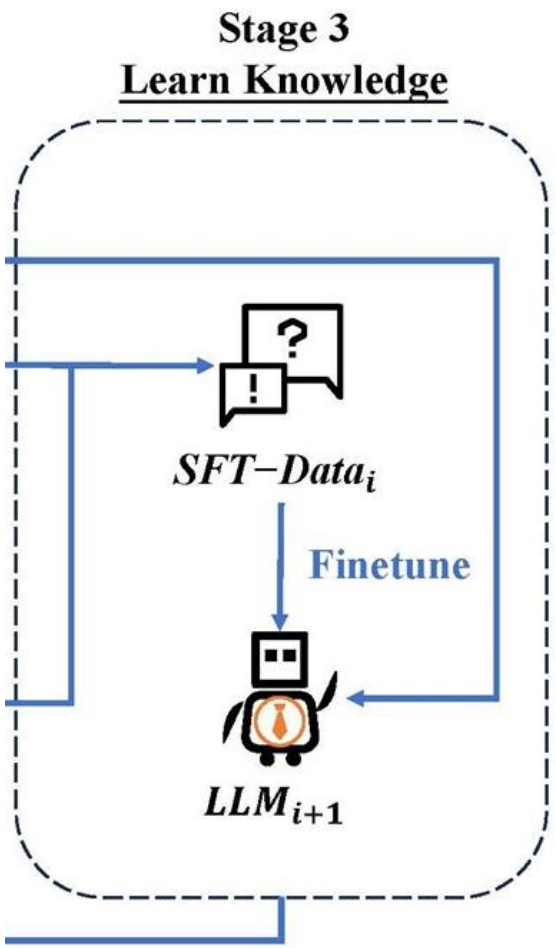
$$s_\theta(A \mid Q) = -\frac{1}{N}\sum_{i=1}^{N} \log P\left(w_i^A \mid Q, w_1^A, \ldots, w_{i-1}^A; \theta\right) \quad (1)$$

$$s_\theta(A) = -\frac{1}{N}\sum_{i=1}^{N} \log P\left(w_i^A \mid w_1^A, w_2^A, \ldots, w_{i-1}^A; \theta\right) \quad (2)$$

$$\mathrm{IFD}_\theta(Q, A) = \frac{s_\theta(A \mid Q)}{s_\theta(A)} \quad (3)$$

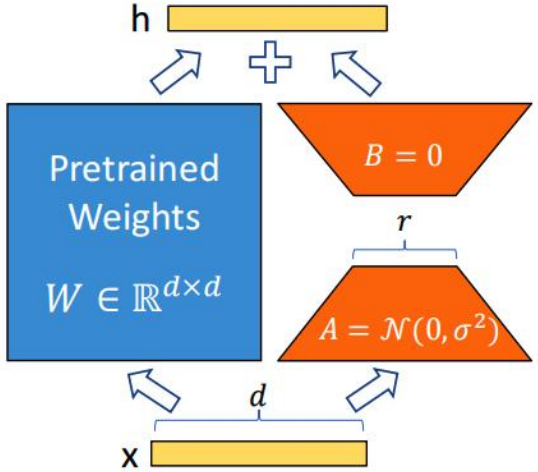
IFD Score Calculation Formula

Learn Knowledge



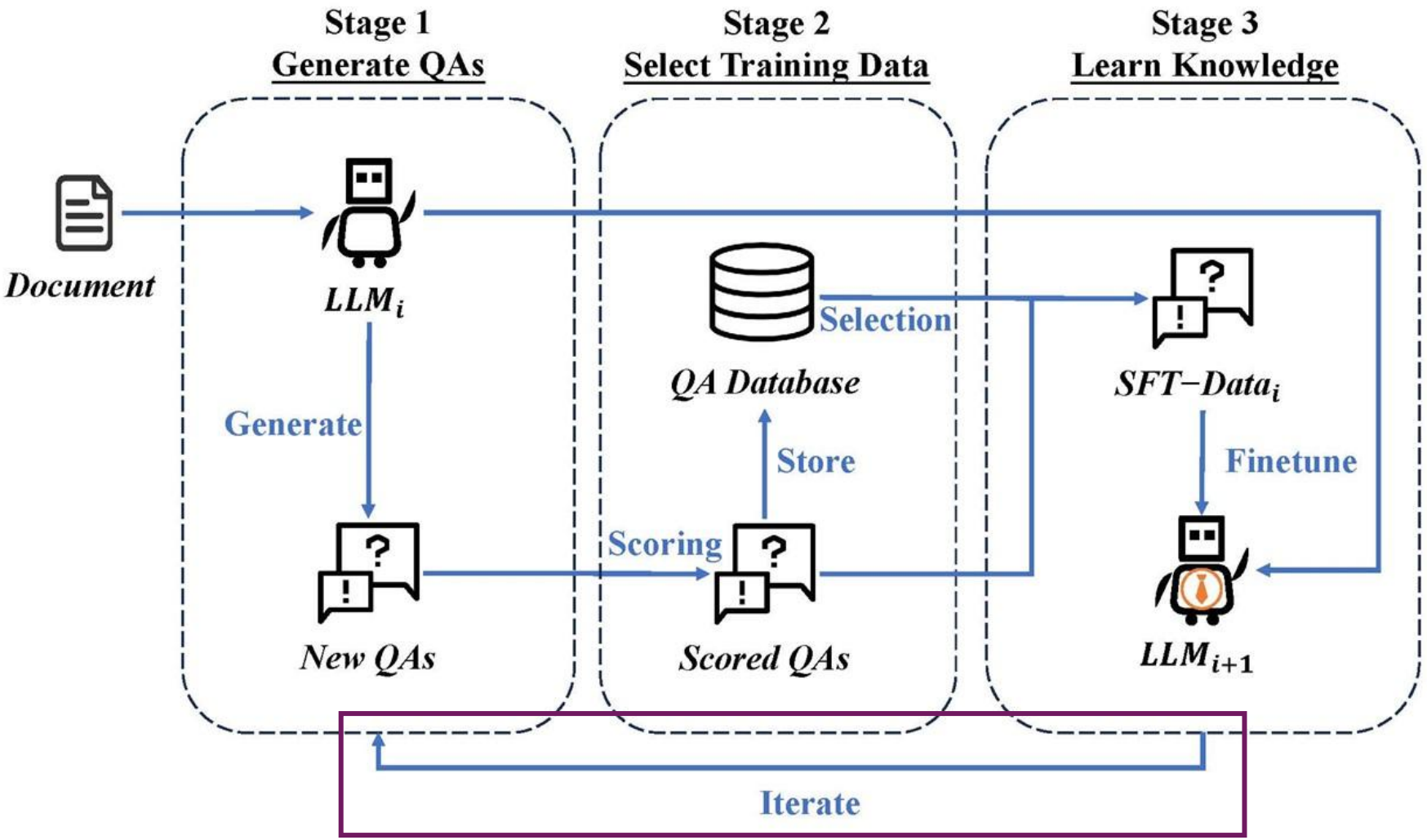
Low−Rank Adaptation of Large Language Models[2] (LoRA)

- Perform low−rank decomposition on the weight matrix
- Accelerate training speed and reduce the demand for computational resources



LoRA

Next Iteration



Outline




Background Design **Evaluation** Conclusion

Baseline

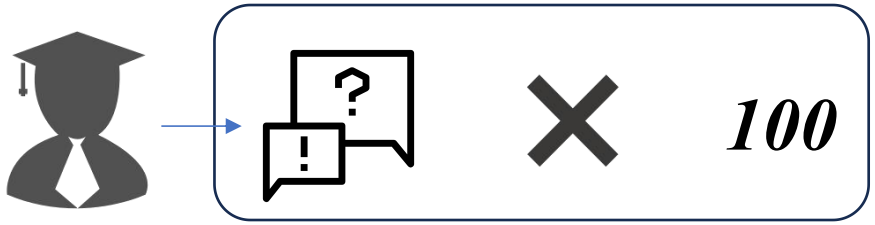
Model	Description
Qwen1.5–7B–HQ (θ_{HQ})	*1. Generate high–quality QA using Qwen1.5–72B–Chat based on the document.* *2. Train Qwen1.5–7B–Chat with the data generated in the first step.* *3. Obtain Qwen1.5–7B–HQ.*
Qwen1.5–7B–Chat	*Original Model*
Qwen1.5–72B–Chat	*Original Model*
GPT–3.5	*Original Model*

Baseline

Dataset

Expert–constructed evaluation set

- 100 question–and–answer items constructed by China Mobile experts.
- The data format is (Question, Answer).



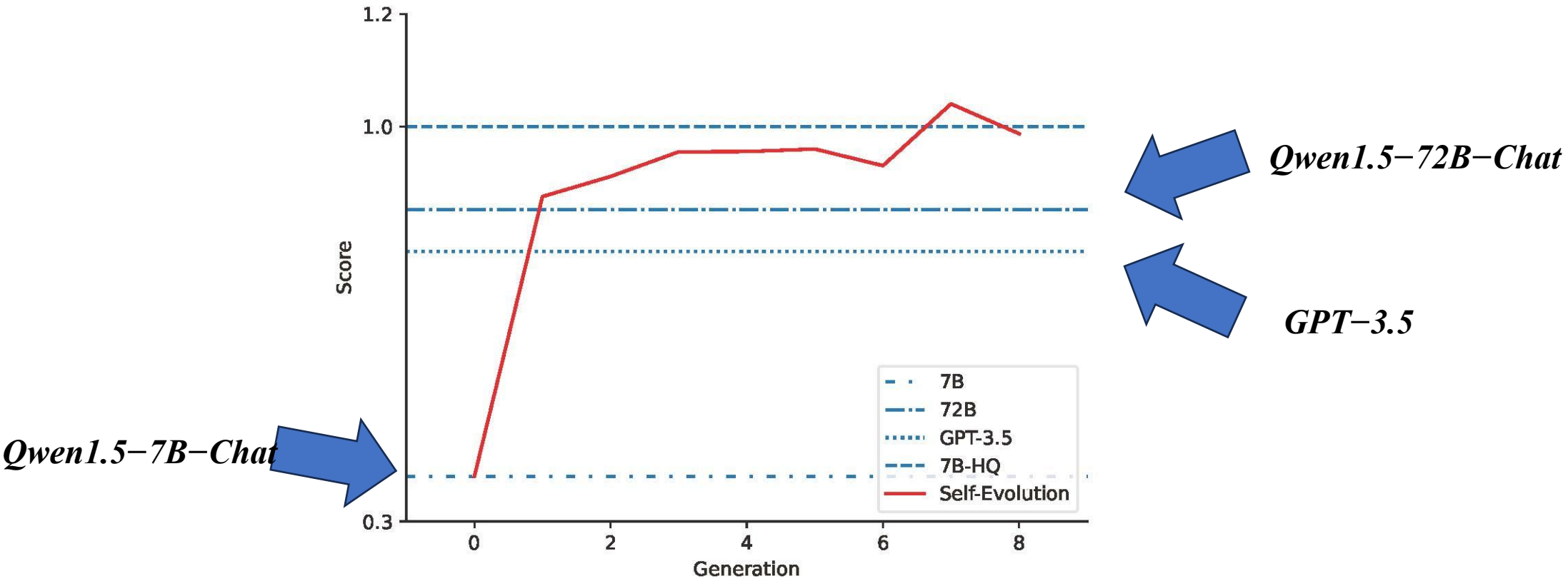
- It includes operation knowledge commonly used within the company, such as **fault description** and **equipment configuration**.

Evaluation method

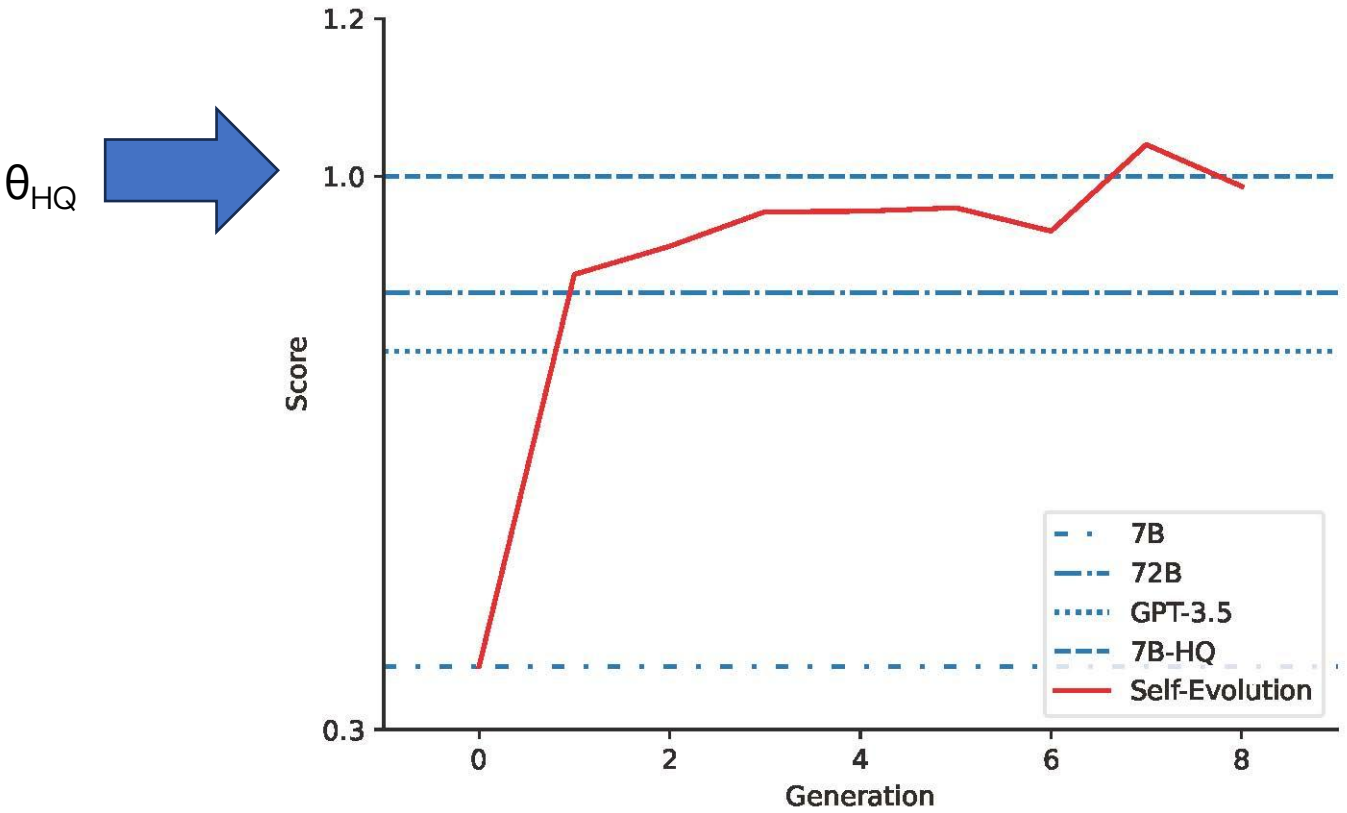
- The model theta's response to the question is compared with the standard answer using the BLEU[3] score.
- For easy demonstration, we uniformly divide by the score of the baseline.

$$Score = \frac{BLEU(\theta)}{BLEU(\theta_{HQ})}$$

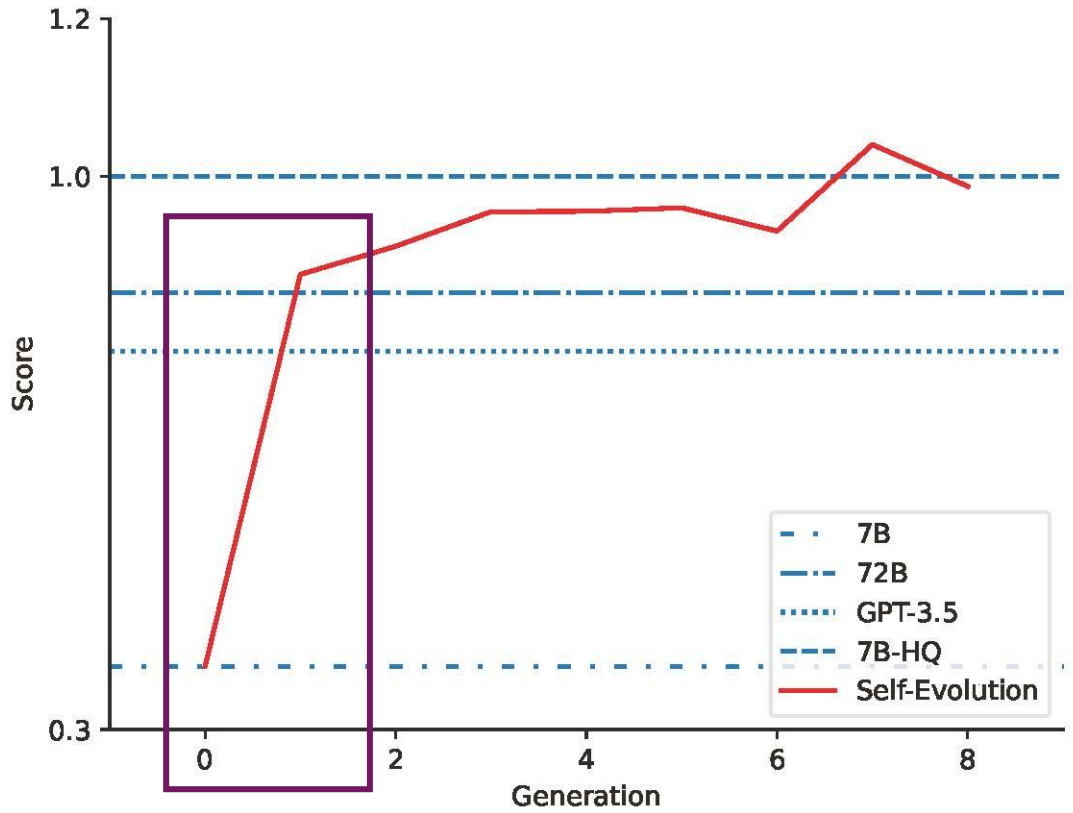
Experimental Results



Experimental Results



Experimental Results

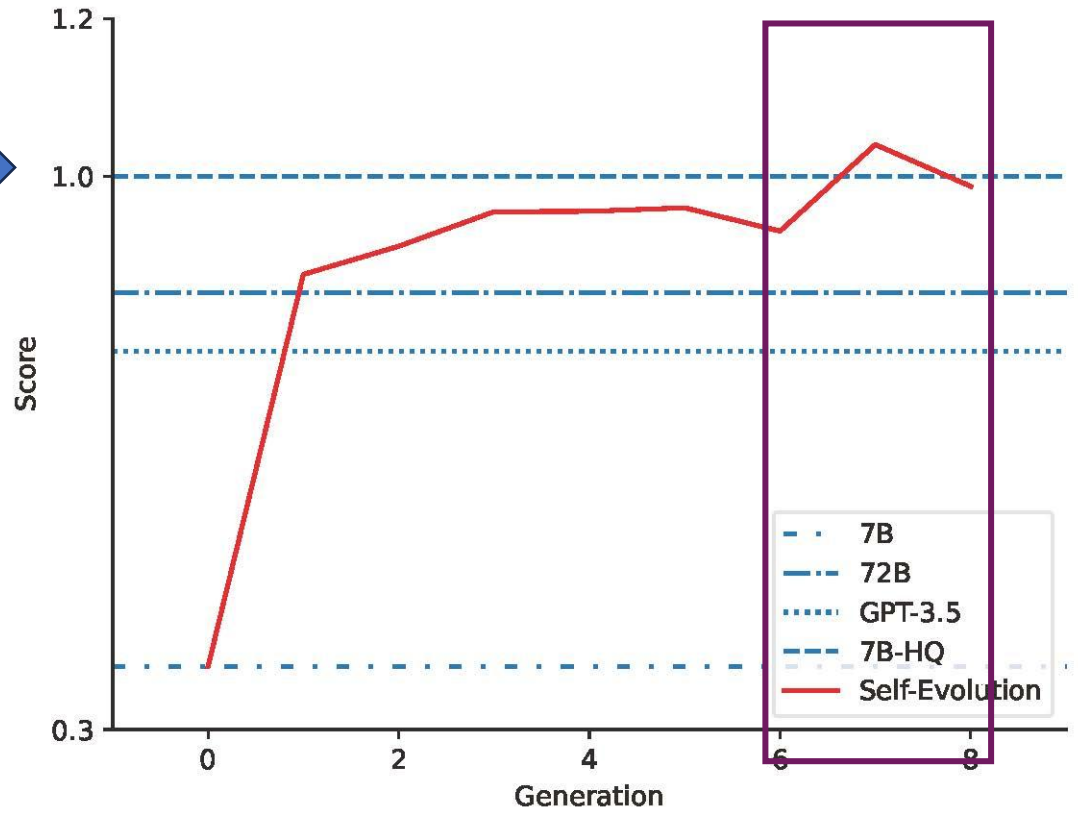


Qwen1.5−72B−Chat

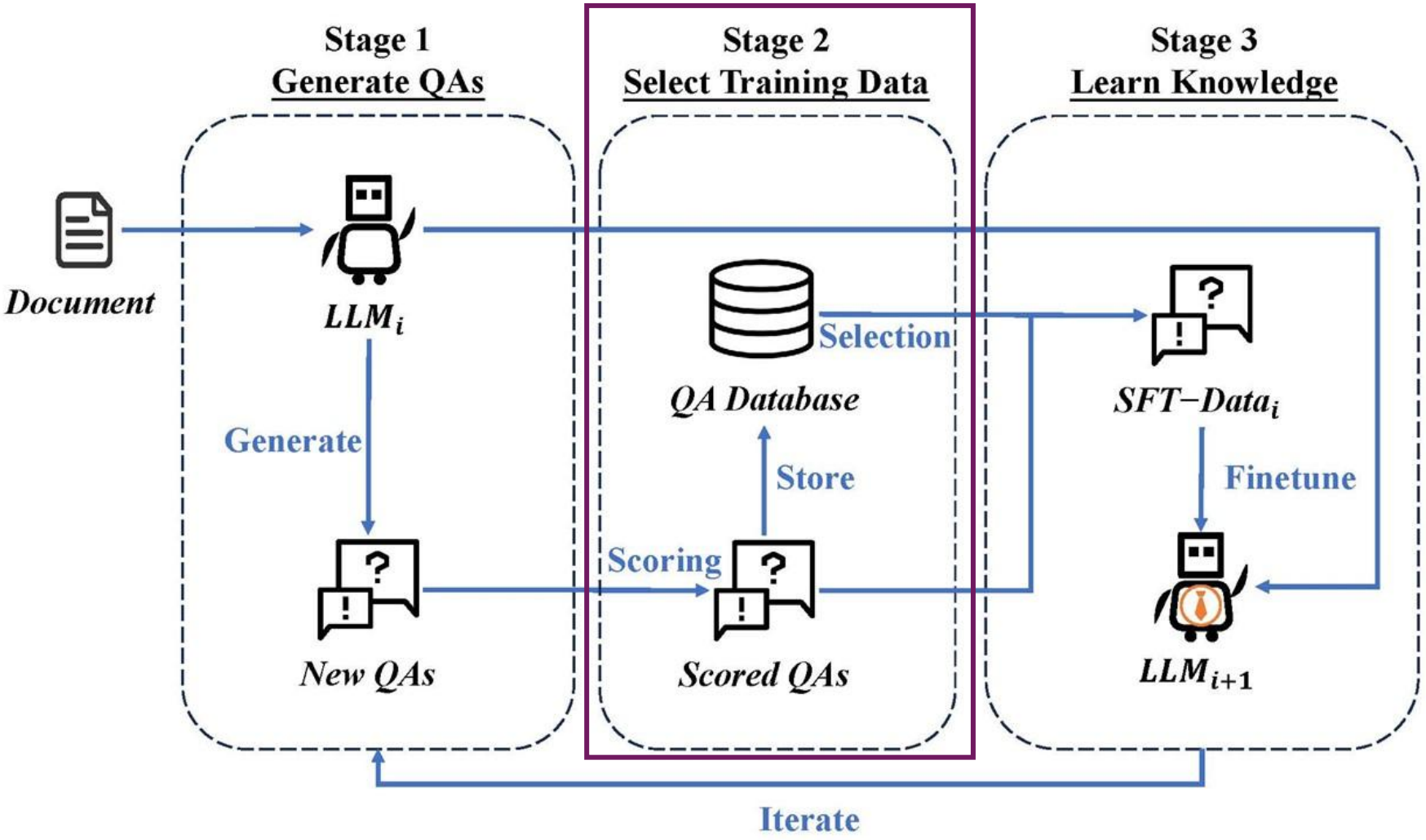
GPT−3.5

Experimental Results

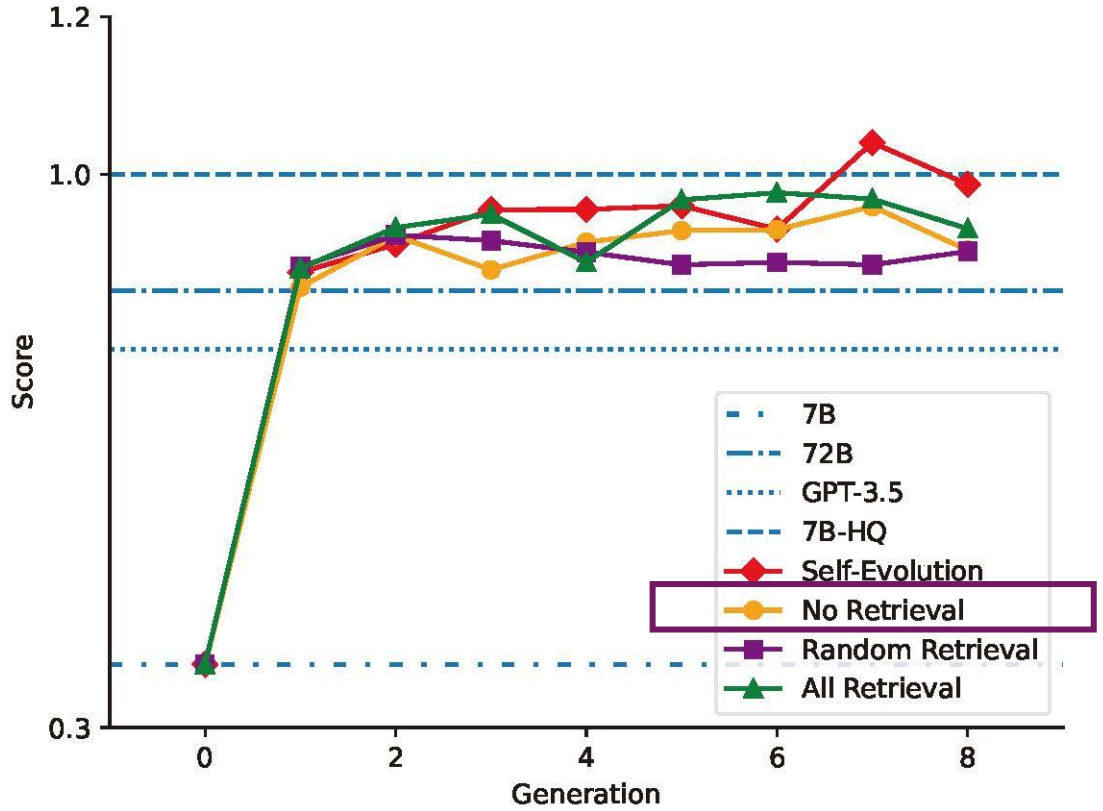
Qwen1.5−7B−HQ



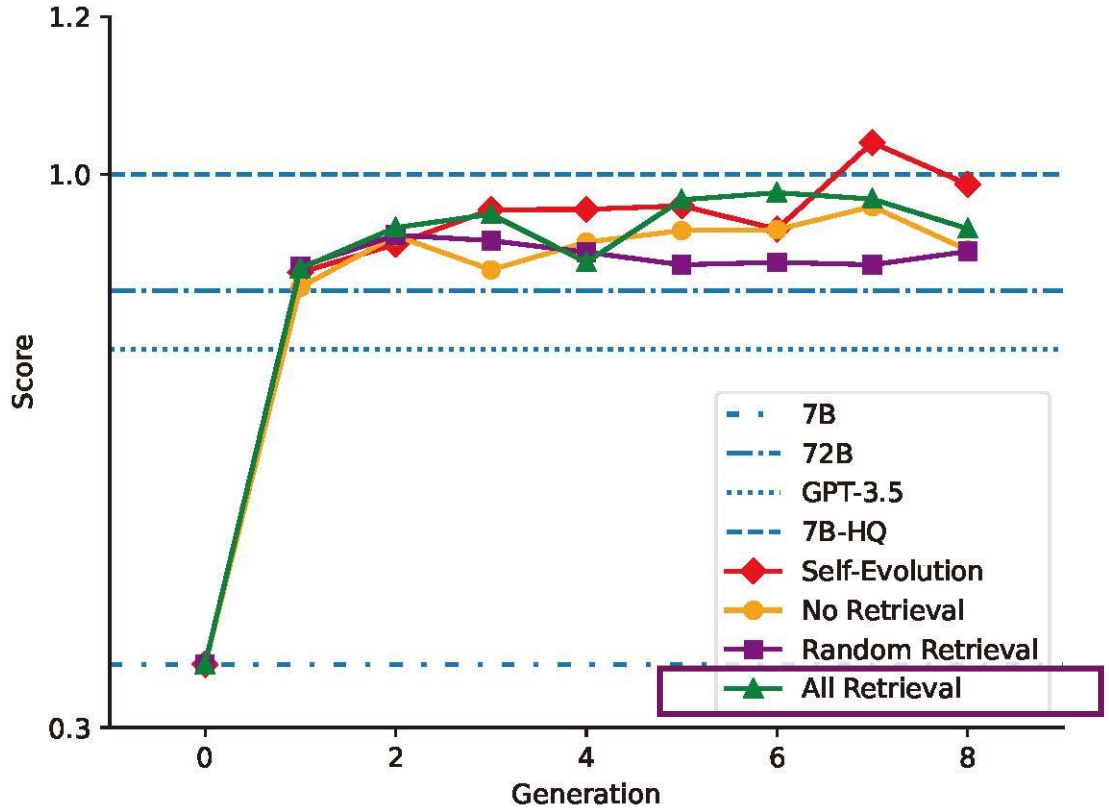
Ablation Study



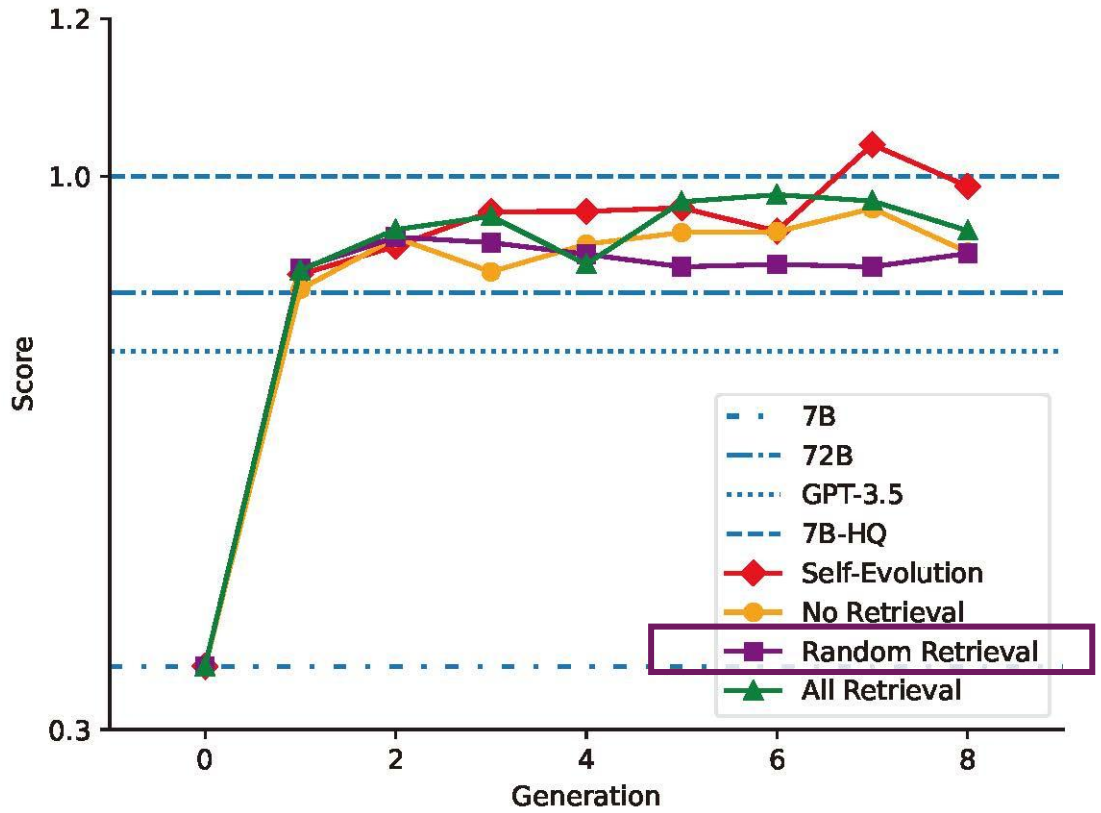
Ablation Study



Ablation Study

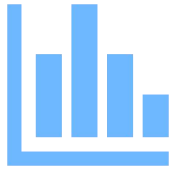


Ablation Study



Outline






Background Design Evaluation **Conclusion**

Conclusion

- The purpose of this work is to realize a lightweight domain–specific Q&A model. To get this, we design a iteratively–SFT framework that make a 7B LLM reaches a 72B model's performance.

- The IFD score is used to select difficult historical QA, by learning multiple times.

- The effectiveness was demonstrated by the sufficient experiments

Reference

[1] M. Li, Y. Zhang, Z. Li, J. Chen, L. Chen, N. Cheng, J. Wang, T. Zhou, and J. Xiao, "From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning," arXiv preprint arXiv:2308.12032, 2023.

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

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

Thank You

sunyongqian@nankai.edu.cn