# Boosting the Potential of Large Language Models with an Intelligent Information Assistant

Yujia Zhou, Zheng Liu, Zhicheng Dou
Tsinghua University
BAAI
Renmin University of China

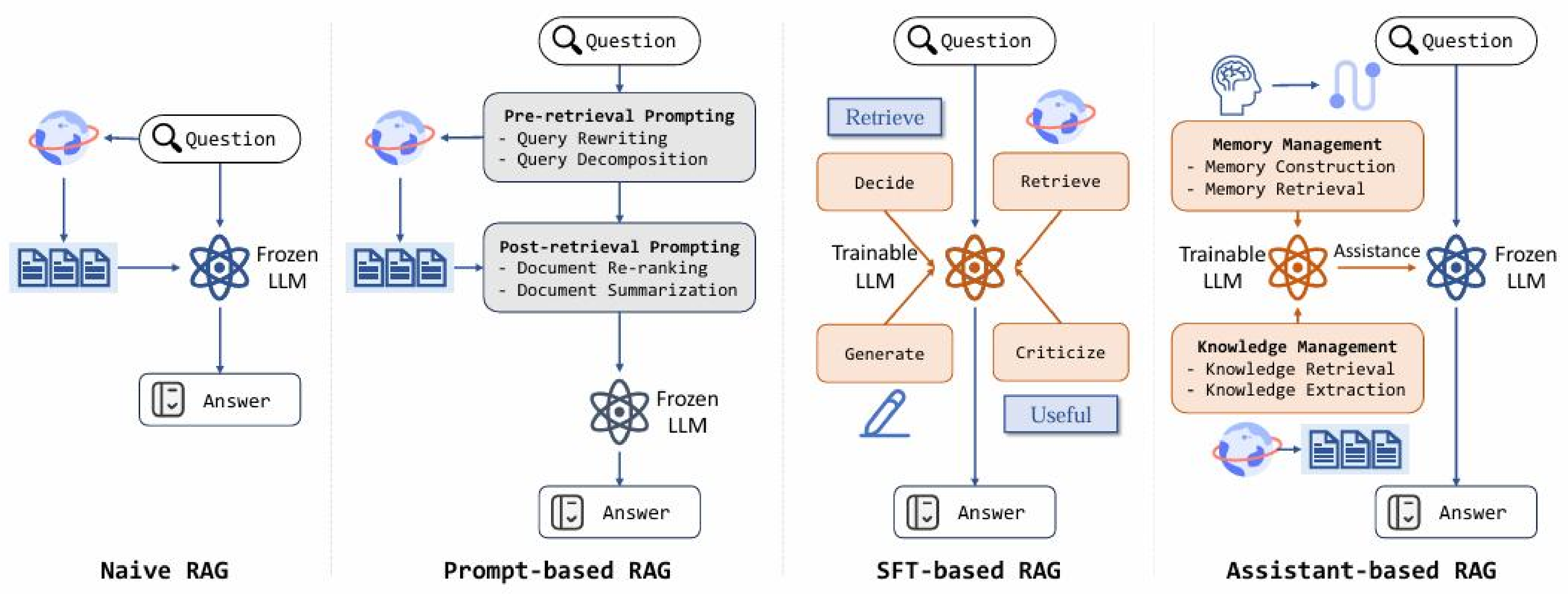
# Introduction

Drawbacks of LLMs

- Hallucination
- Outdated information
- Low efficiency in parameterizing knowledge
- Lack of in-depth knowledge in specialized domains
- Weak inferential capabilities

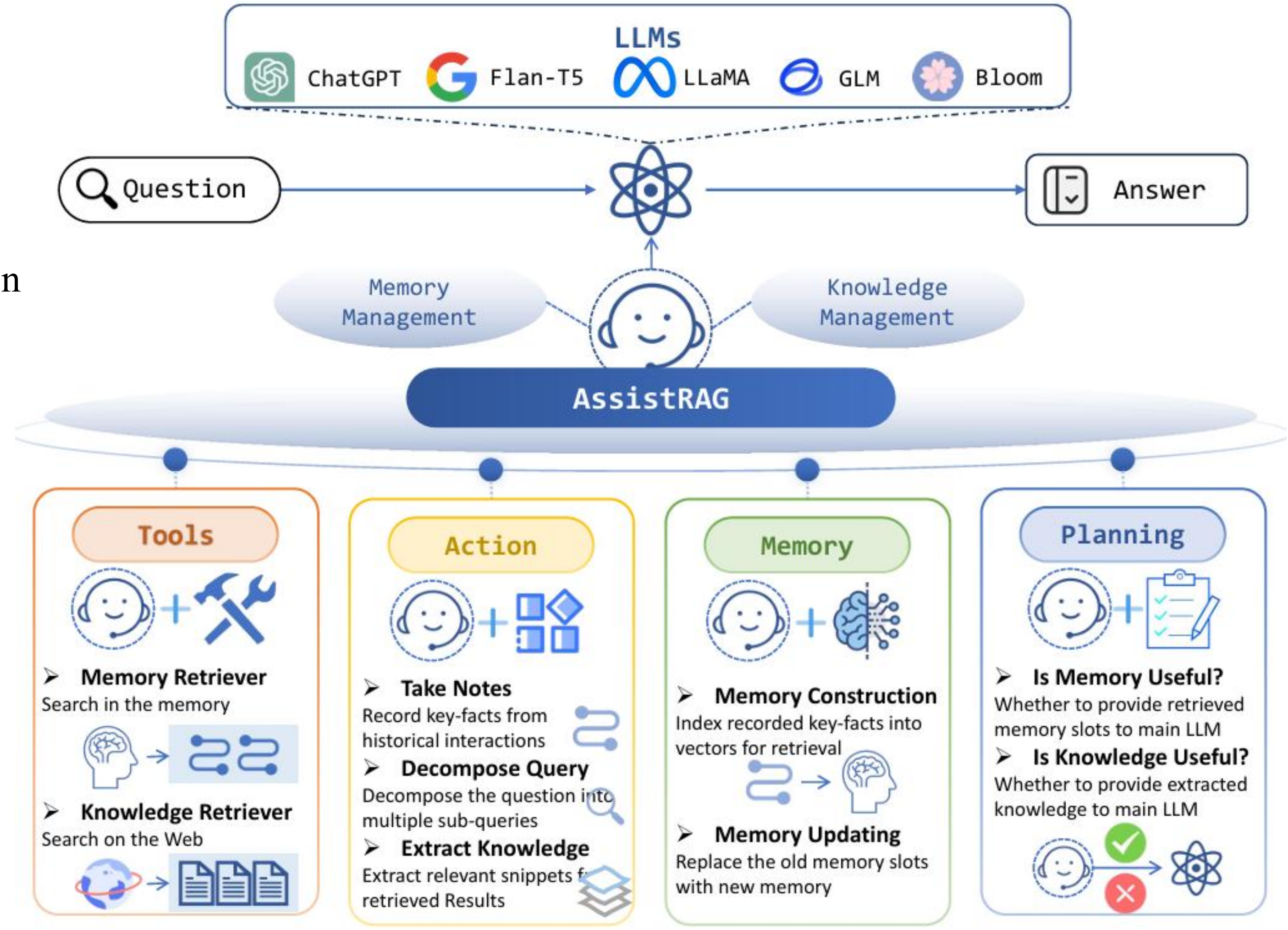
Retrieval-augemented generation

# Introduction

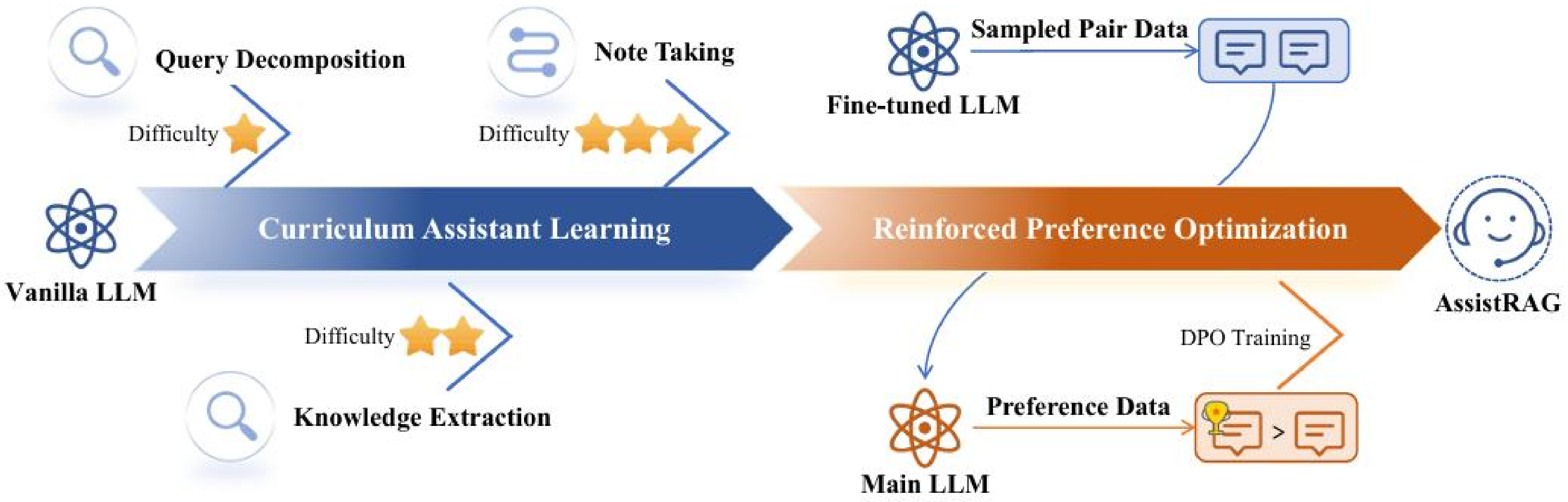


# AssistRAG Framework

- **Tool Usage:** Retrieving relevant information from internal memory and external knowledge bases.
- **Action Execution:** Reasoning, analyzing information need, and extracting knowledge.
- **Memory Building:** Recording essential knowledge and reasoning patterns from past interactions.
- **Plan Specification:** Determining the necessity of assistance during answer generation.

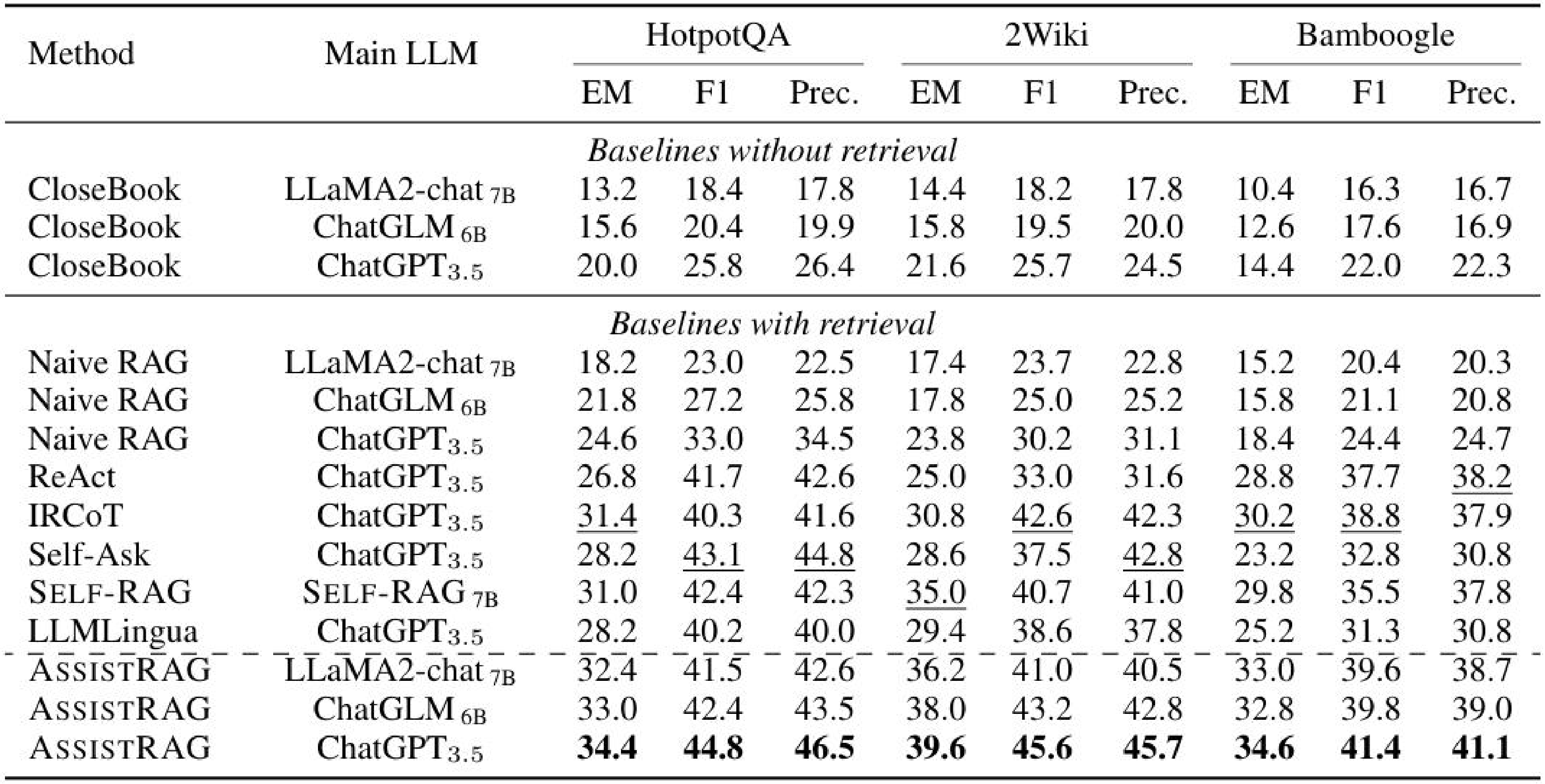

LLMs
ChatGPT
Flan-T5
LLaMA
GLM
Bloom
Question
Answer
Memory Management
Knowledge Management
AssistRAG
Tools
Memory Retriever
Search in the memory
Knowledge Retriever
Search on the Web
Action
Take Notes
Record key-facts from historical interactions
Decompose Query
Decompose the question into multiple sub-queries
Extract Knowledge
Extract relevant snippets f retrieved Results
Memory
Memory Construction
Index recorded key-facts into vectors for retrieval
Memory Updating
Replace the old memory slots with new memory
Planning
Is Memory Useful?
Whether to provide retrieved memory slots to main LLM
Is Knowledge Useful?
Whether to provide extracted knowledge to main LLM

# AssistRAG Training



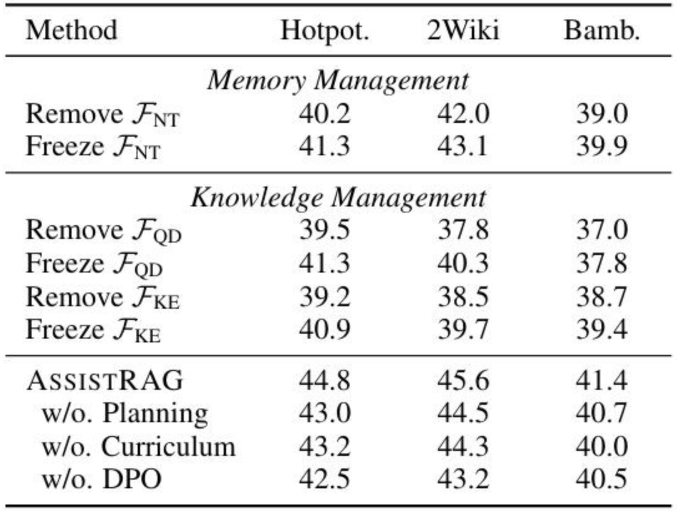
- **Curriculum Assistant Learning** enhances the assistant's capabilities in note-taking, question decomposition, and knowledge extraction through progressively complex tasks.
- **Reinforced Preference Optimization** uses reinforcement learning to tailor the assistant's feedback to the main LLM's specific needs, optimizing knowledge extraction based on feedback from the main LLM.

# Introduction

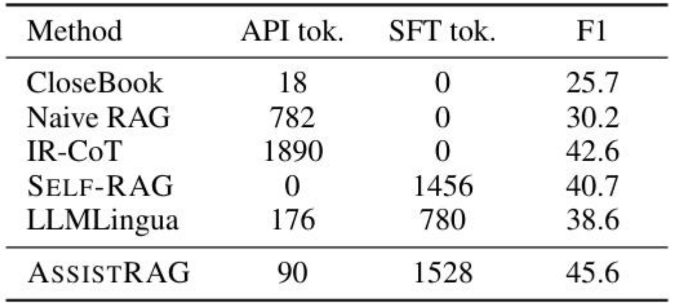


| Method | Main LLM | HotpotQA |  |  | 2Wiki |  |  | Bamboogle |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  |  | EM | F1 | Prec. | EM | F1 | Prec. | EM | F1 | Prec. |
| *Baselines without retrieval* |  |  |  |  |  |  |  |  |  |  |
| CloseBook | LLaMA2-chat$_{7B}$ | 13.2 | 18.4 | 17.8 | 14.4 | 18.2 | 17.8 | 10.4 | 16.3 | 16.7 |
| CloseBook | ChatGLM$_{6B}$ | 15.6 | 20.4 | 19.9 | 15.8 | 19.5 | 20.0 | 12.6 | 17.6 | 16.9 |
| CloseBook | ChatGPT$_{3.5}$ | 20.0 | 25.8 | 26.4 | 21.6 | 25.7 | 24.5 | 14.4 | 22.0 | 22.3 |
| *Baselines with retrieval* |  |  |  |  |  |  |  |  |  |  |
| Naive RAG | LLaMA2-chat$_{7B}$ | 18.2 | 23.0 | 22.5 | 17.4 | 23.7 | 22.8 | 15.2 | 20.4 | 20.3 |
| Naive RAG | ChatGLM$_{6B}$ | 21.8 | 27.2 | 25.8 | 17.8 | 25.0 | 25.2 | 15.8 | 21.1 | 20.8 |
| Naive RAG | ChatGPT$_{3.5}$ | 24.6 | 33.0 | 34.5 | 23.8 | 30.2 | 31.1 | 18.4 | 24.4 | 24.7 |
| ReAct | ChatGPT$_{3.5}$ | 26.8 | 41.7 | 42.6 | 25.0 | 33.0 | 31.6 | 28.8 | 37.7 | 38.2 |
| IRCoT | ChatGPT$_{3.5}$ | 31.4 | 40.3 | 41.6 | 30.8 | 42.6 | 42.3 | 30.2 | 38.8 | 37.9 |
| Self-Ask | ChatGPT$_{3.5}$ | 28.2 | 43.1 | 44.8 | 28.6 | 37.5 | 42.8 | 23.2 | 32.8 | 30.8 |
| Self-RAG | Self-RAG$_{7B}$ | 31.0 | 42.4 | 42.3 | 35.0 | 40.7 | 41.0 | 29.8 | 35.5 | 37.8 |
| LLMLingua | ChatGPT$_{3.5}$ | 28.2 | 40.2 | 40.0 | 29.4 | 38.6 | 37.8 | 25.2 | 31.3 | 30.8 |
| AssistRAG | LLaMA2-chat$_{7B}$ | 32.4 | 41.5 | 42.6 | 36.2 | 41.0 | 40.5 | 33.0 | 39.6 | 38.7 |
| AssistRAG | ChatGLM$_{6B}$ | 33.0 | 42.4 | 43.5 | 38.0 | 43.2 | 42.8 | 32.8 | 39.8 | 39.0 |
| AssistRAG | ChatGPT$_{3.5}$ | **34.4** | **44.8** | **46.5** | **39.6** | **45.6** | **45.7** | **34.6** | **41.4** | **41.1** |

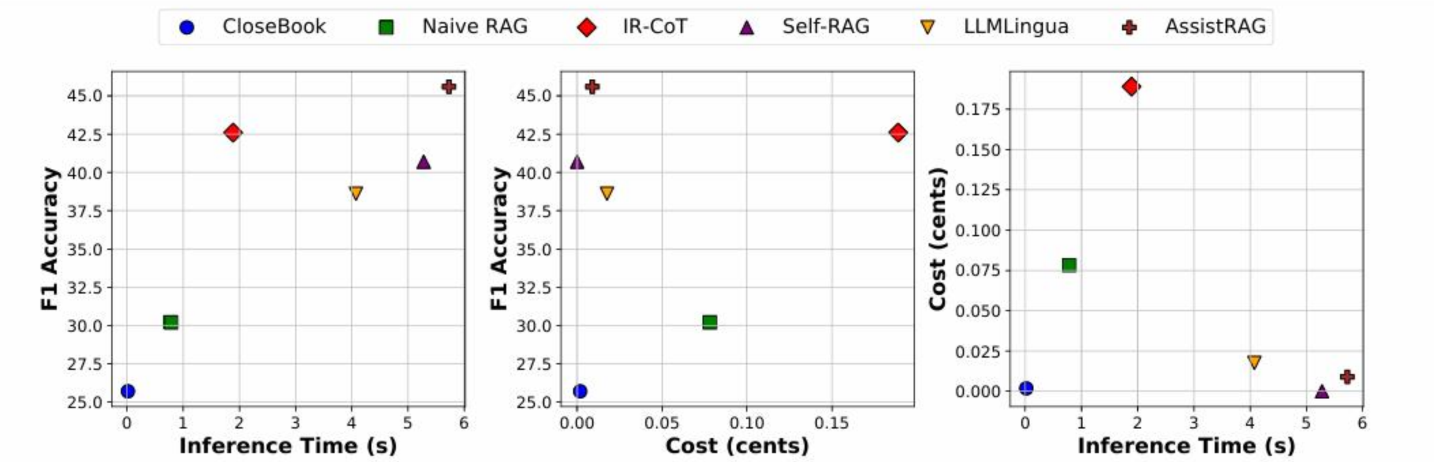
# Introduction

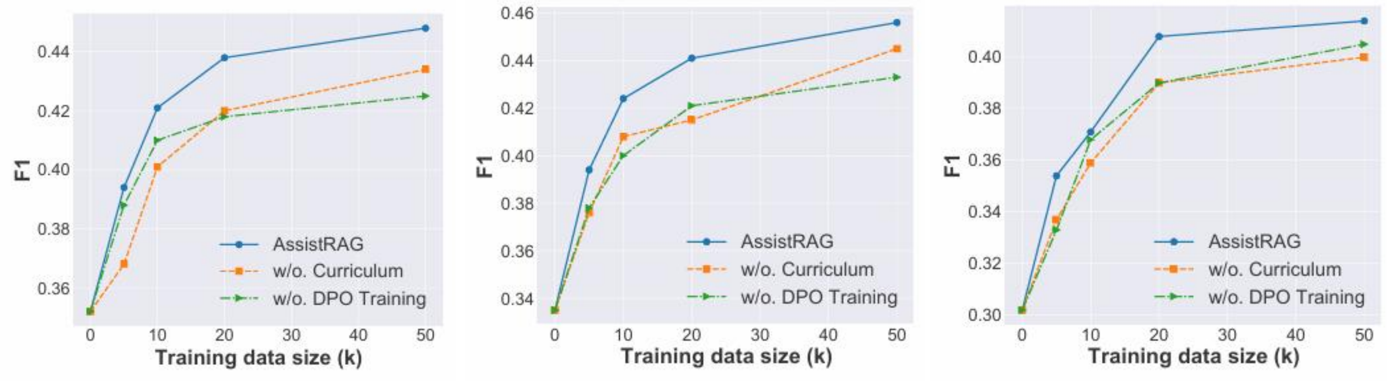


| Method | Hotpot. | 2Wiki | Bamb. |
|---|---|---|---|
| *Memory Management* | | | |
| Remove $\mathcal{F}_{\text{NT}}$ | 40.2 | 42.0 | 39.0 |
| Freeze $\mathcal{F}_{\text{NT}}$ | 41.3 | 43.1 | 39.9 |
| *Knowledge Management* | | | |
| Remove $\mathcal{F}_{\text{QD}}$ | 39.5 | 37.8 | 37.0 |
| Freeze $\mathcal{F}_{\text{QD}}$ | 41.3 | 40.3 | 37.8 |
| Remove $\mathcal{F}_{\text{KE}}$ | 39.2 | 38.5 | 38.7 |
| Freeze $\mathcal{F}_{\text{KE}}$ | 40.9 | 39.7 | 39.4 |
| ASSISTRAG | 44.8 | 45.6 | 41.4 |
| w/o. Planning | 43.0 | 44.5 | 40.7 |
| w/o. Curriculum | 43.2 | 44.3 | 40.0 |
| w/o. DPO | 42.5 | 43.2 | 40.5 |



| Method | API tok. | SFT tok. | F1 |
|---|---|---|---|
| CloseBook | 18 | 0 | 25.7 |
| Naive RAG | 782 | 0 | 30.2 |
| IR-CoT | 1890 | 0 | 42.6 |
| SELF-RAG | 0 | 1456 | 40.7 |
| LLMLingua | 176 | 780 | 38.6 |
| ASSISTRAG | 90 | 1528 | 45.6 |





(a) HotpotQA

(b) 2WikiMultiHopQA

(c) Bamboogle