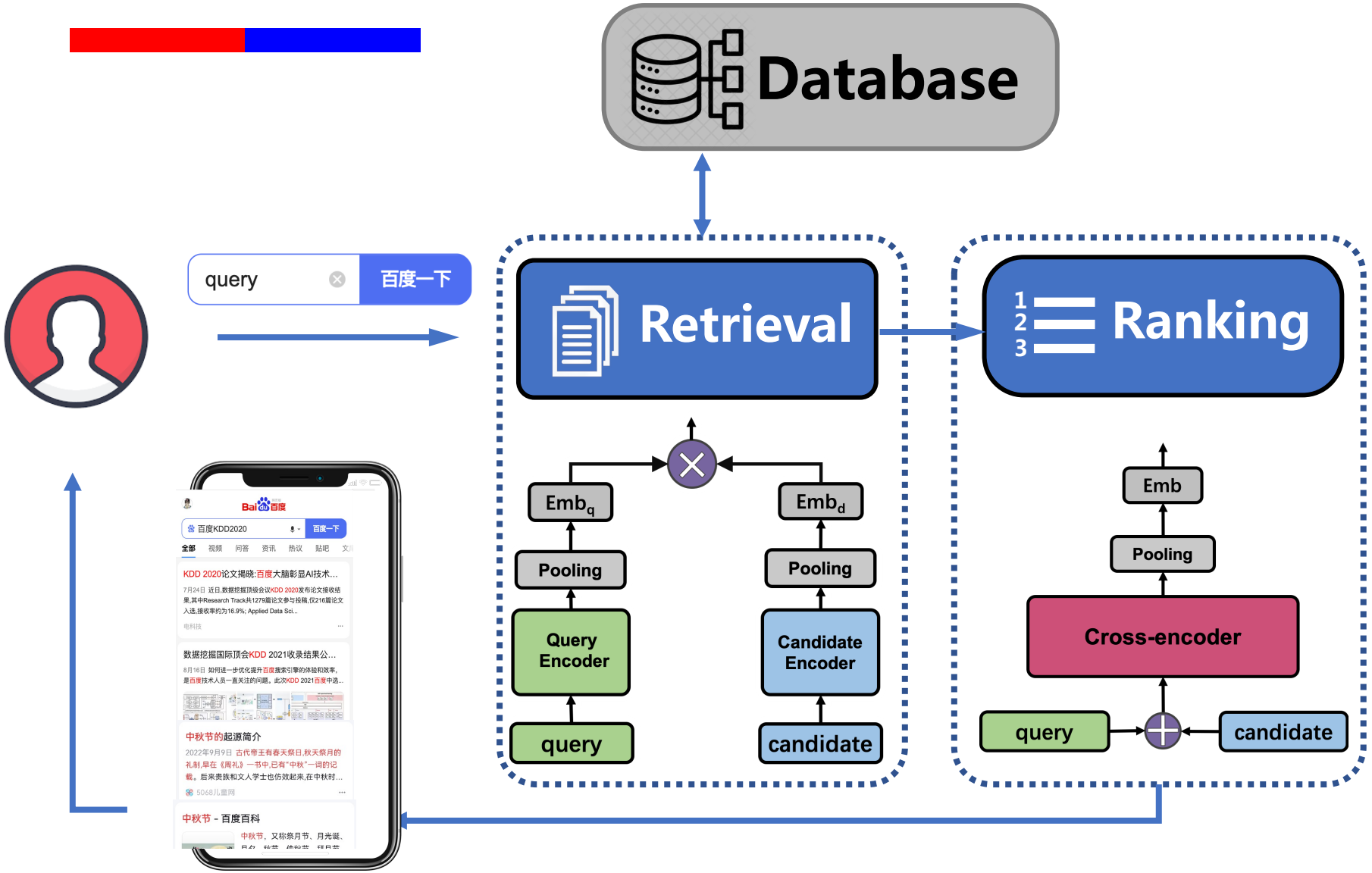
Retrieval-Augmented Generation via LLMs in Web Search

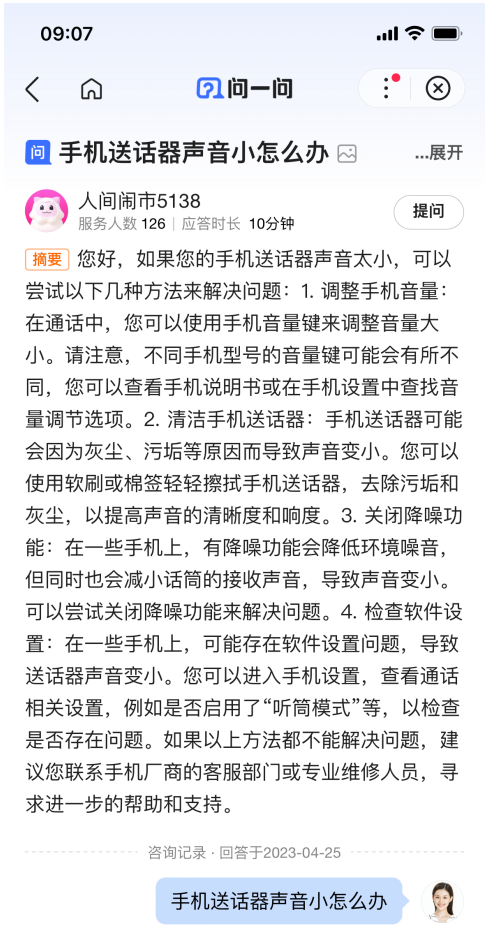
Shuaiqiang Wang
Baidu Search
http://searchscience.baidu.com

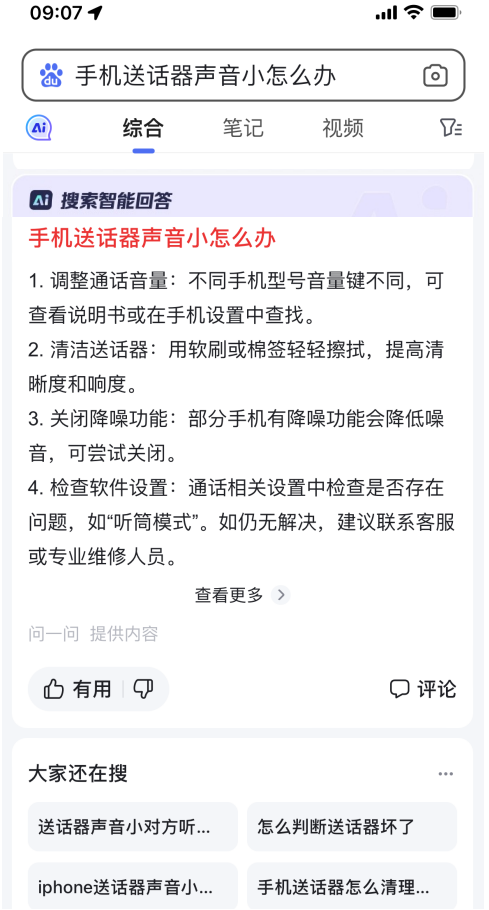
Conventional Search Engine

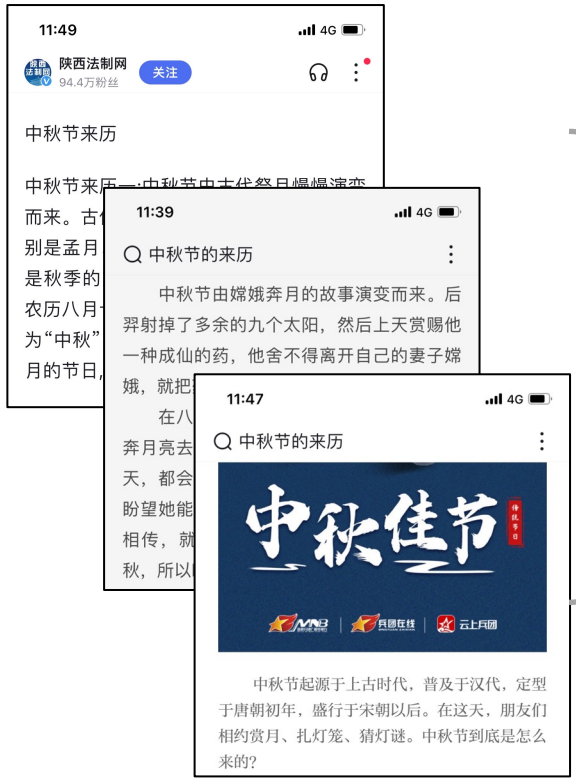


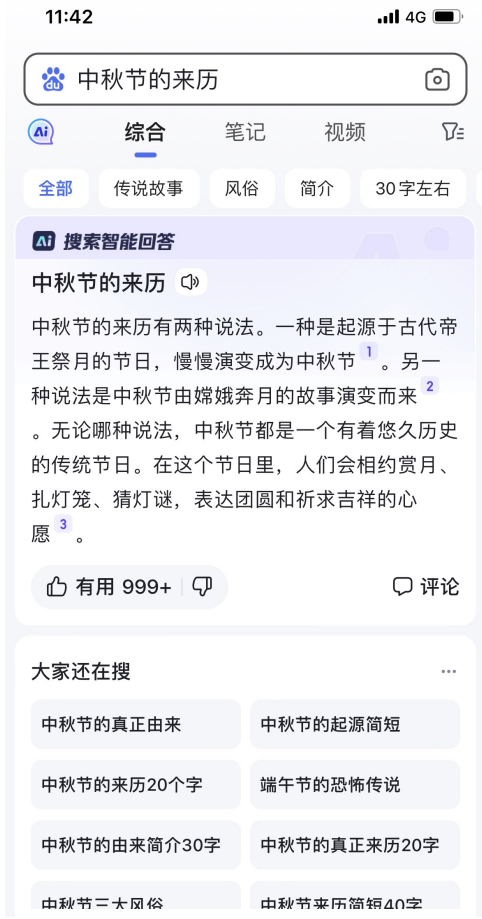
1. Zou et al. Pre-trained Language Model based Ranking in Baidu Search. In KDD 2021.
2. Liu et al. Pre-trained Language Model for Web-scale Retrieval in Baidu Search. In KDD 2021.
3. Ye et al. Fast Semantic Matching via Flexible Contextualized Interaction. In WSDM 2022.
4. Dong et al. Incorporating Explicit Knowledge in Pre-trained Language Models for Passage Re-ranking. In SIGIR 2022.
5. Zou et al. A Large Scale Search Dataset for Unbiased Learning to Rank. In NeurIPS 2022.

Retrieval-Augmented Generation in Baidu

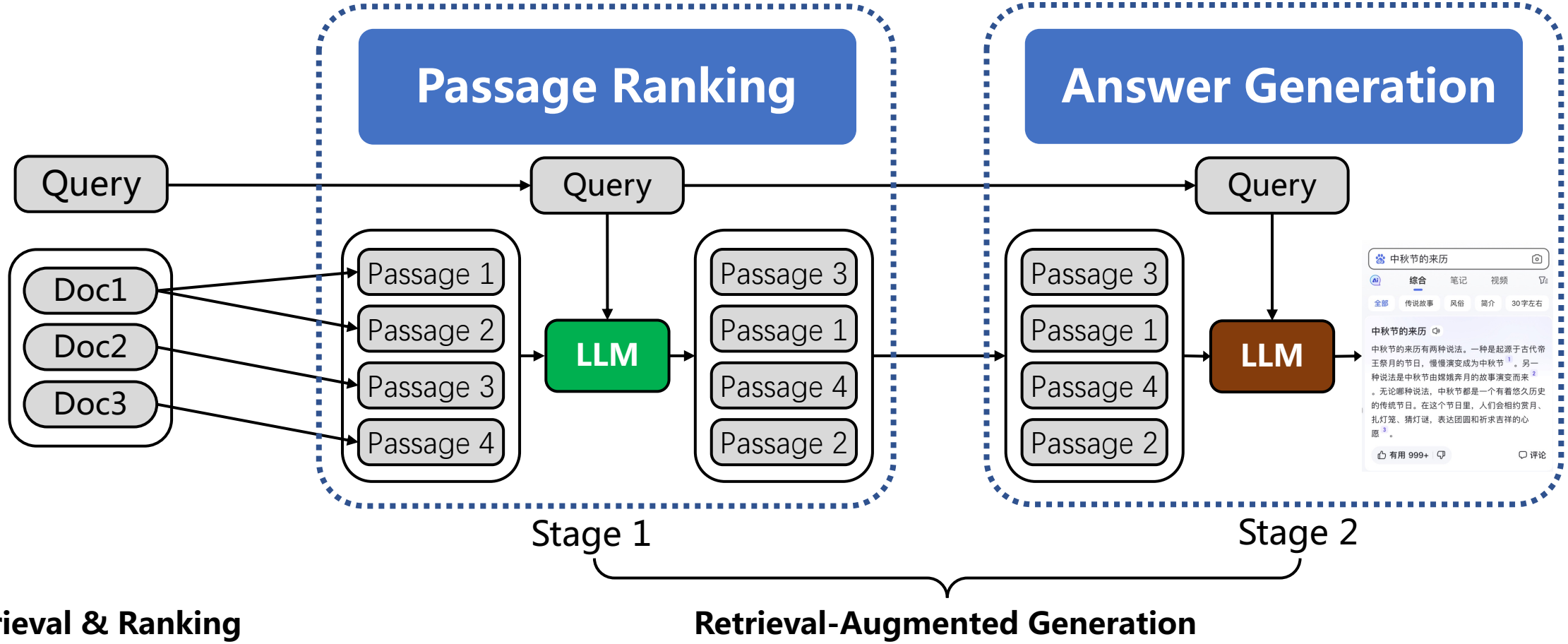




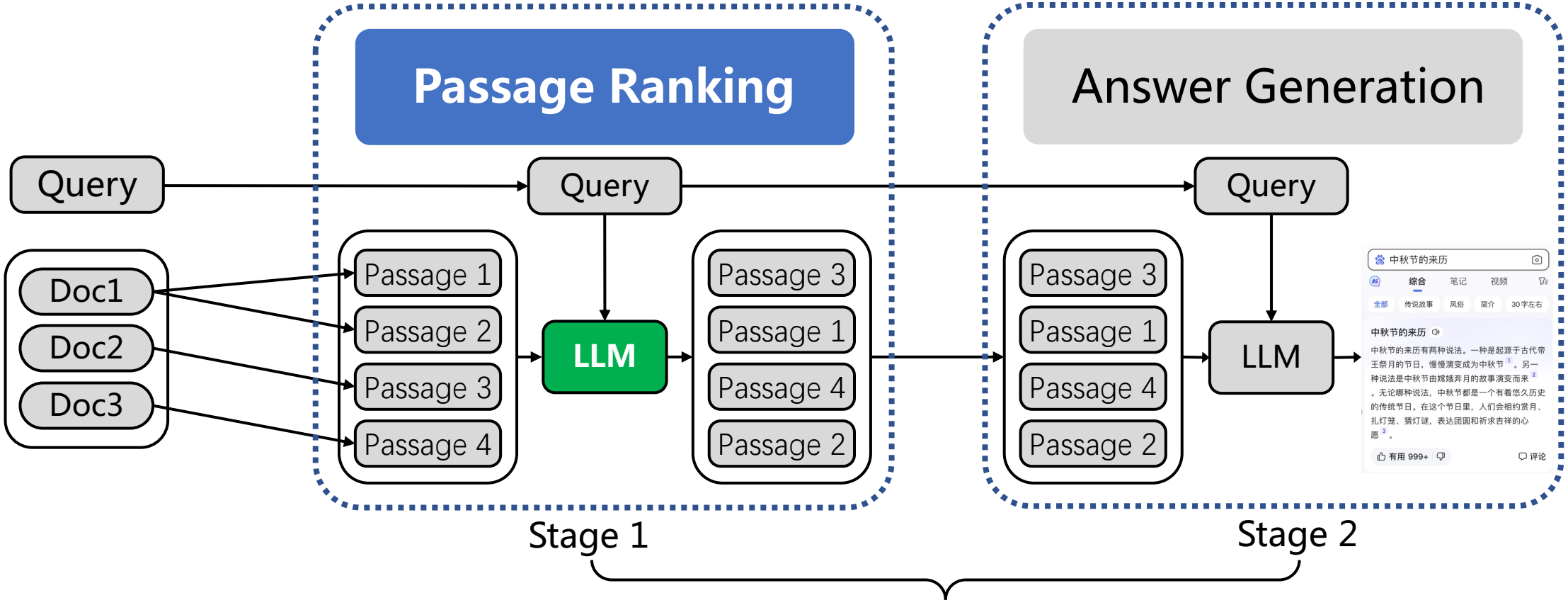




Retrieval-Augmented Generation in Baidu

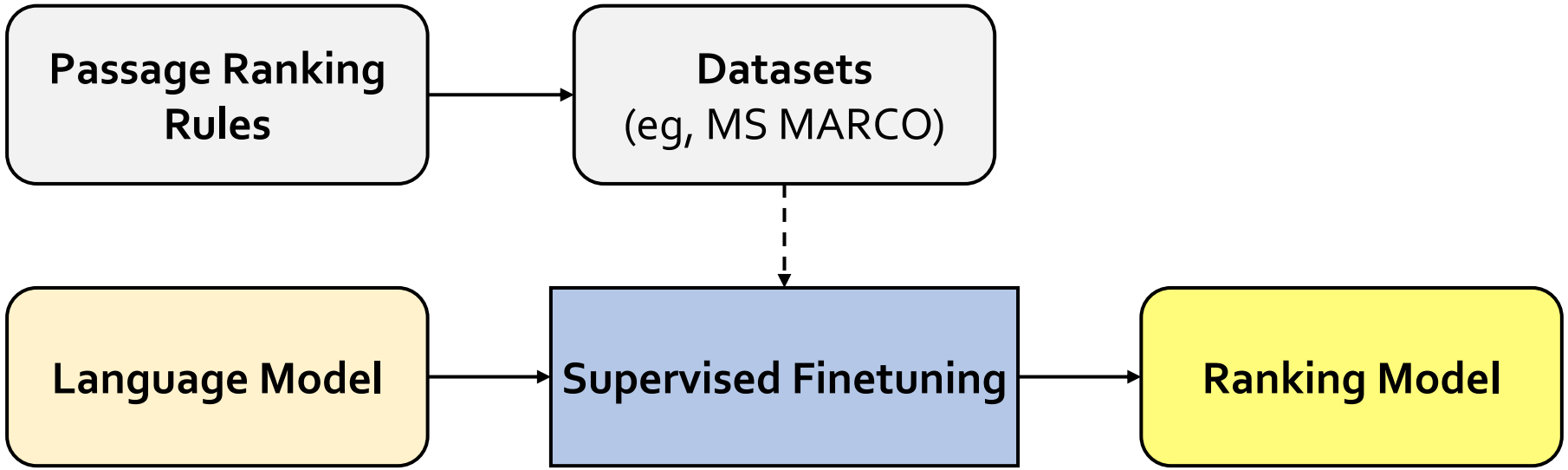


Retrieval-Augmented Generation in Baidu



Passage Ranking

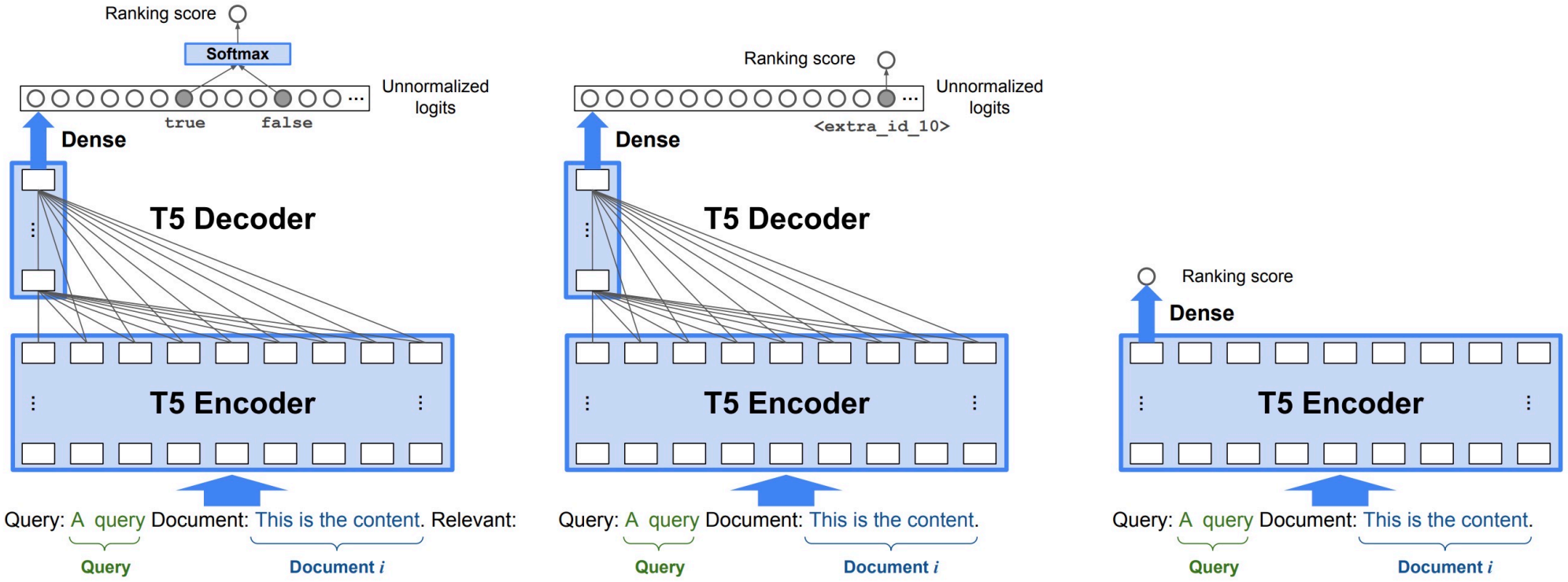
- Fine-tuning based passage ranking



[1] Nogueira et al. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *EMNLP (Findings) 2020*.
[2] Zhuang et al. RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses. In *SIGIR 2023*.

Passage Ranking by LLMs

- Fine-tuning based ranking



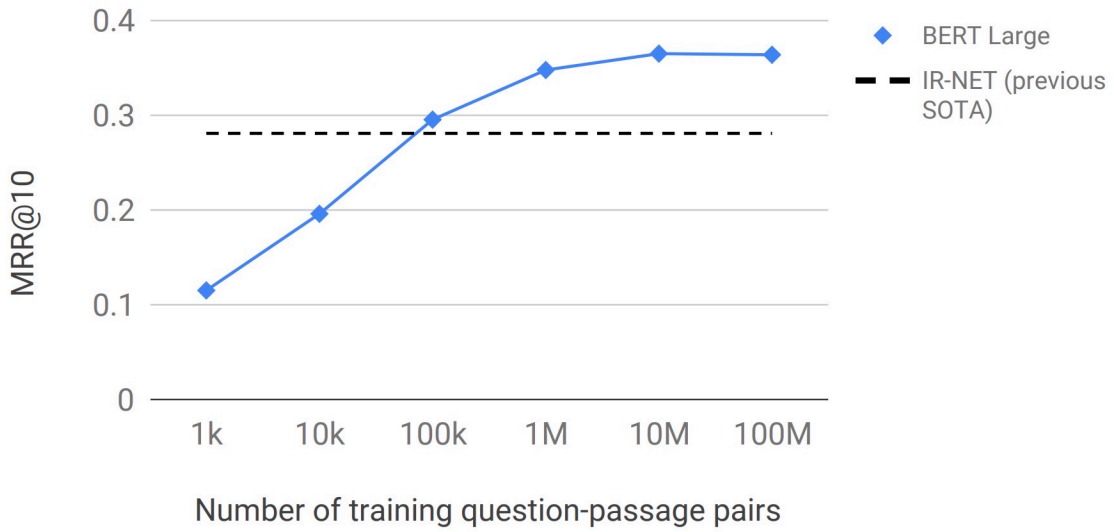
(a) monoT5 (Nogueira et al., 2020) (b) RankT5 (Encoder-Decoder) (c) RankT5 (Encoder-Only)

[1] Nogueira et al. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *EMNLP (Findings) 2020*.
[2] Zhuang et al. RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses. In *SIGIR 2023*.

Passage Ranking by LLMs

- Challenges of fine-tuning based ranking

Significant human effort



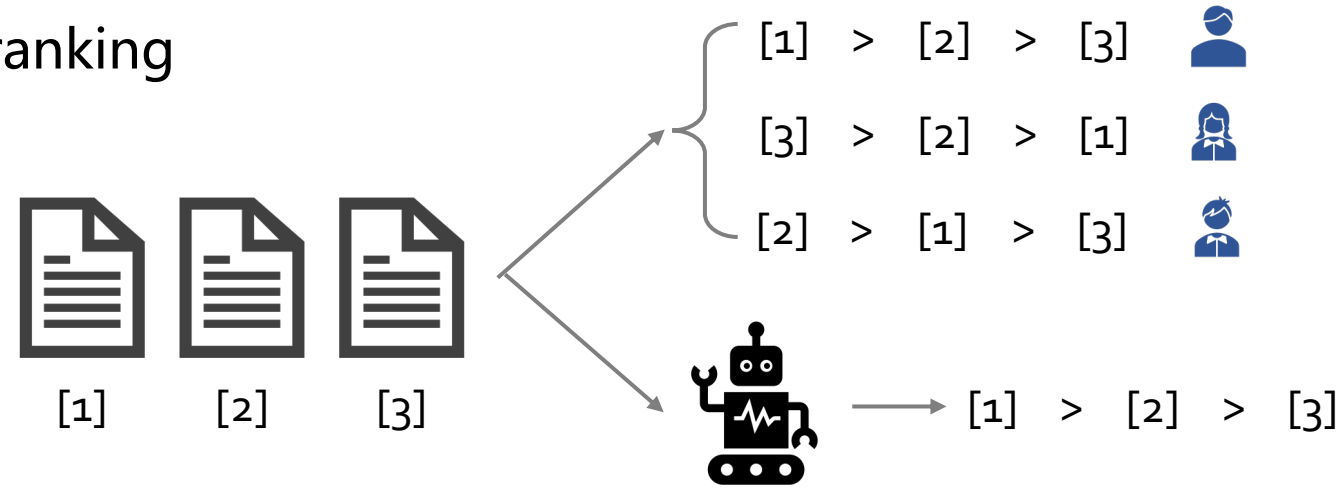
* Picture from: Nogueira, and Cho. "Passage Re-ranking with BERT" *arXiv preprint arXiv: 1901.04085* (2019).

Passage Ranking by LLMs

- Challenges of fine-tuning based ranking

Significant human effort

Inconsistent annotation



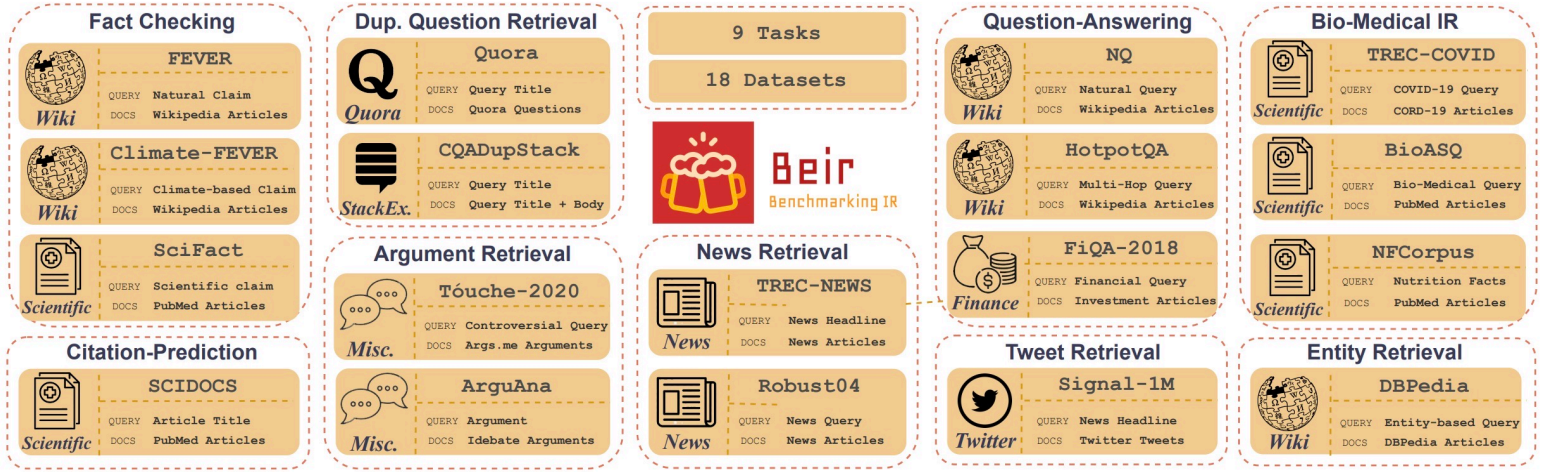
Passage Ranking by LLMs

- Challenges of fine-tuning based ranking

Significant human effort

Inconsistent annotation

Weak generalizability

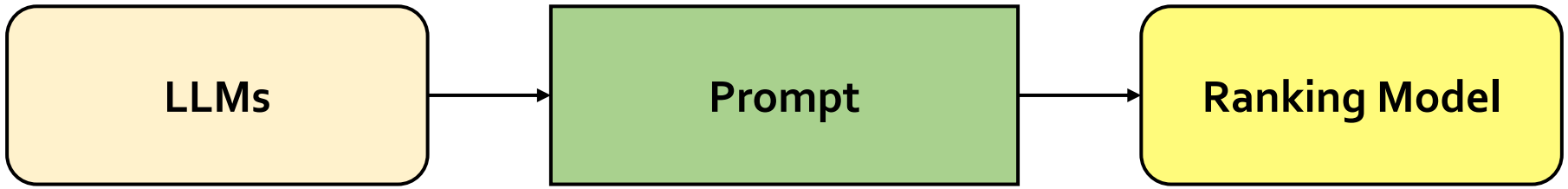


Diverse Ranking Tasks

* Picture from: Thakur. " BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models" *arXiv preprint arXiv: 2104.08663* (2021).

Passage Ranking by LLMs

- LLM-based passage ranking



Ranking with prompt by LLMs

Sun et al. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *EMNLP 2023*.

Passage Ranking by LLMs

- LLM-based passage ranking
  - **Permutation Generation**

A **chat-format prompt** was found to further improve performance.

system:
You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

user:
I will provide you with `{{num}}` passages, each indicated by number identifier []. Rank them based on their relevance to query: `{{query}}`.

assistant:
Okay, please provide the passages.

user:
[1] `{{passage_1}}`

assistant:
Received passage [1]

user:
[2] `{{passage_2}}`

assistant:
Received passage [2]

(more passages) ...

user
Search Query: `{{query}}`.
Rank the `{{num}}` passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers, and the most relevant passages should be listed first, and the output format should be [] > [], e.g., [1] > [2]. Only response the ranking results, do not say any word or explain.

Passage Ranking by LLMs

- LLM-based passage ranking
  - **Permutation Generation**

A **chat-format prompt** was found to further improve performance.

Rank more passages may exceed the **maximum length** limit of the Transformer

system:
You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.

user:
I will provide you with {{num}} passages, each indicated by number identifier []. Rank them based on their relevance to query: {{query}}.

assistant:
Okay, please provide the passages.

user:
[1] {{passage_1}}

assistant:
Received passage [1]

user:
[2] {{passage_2}}

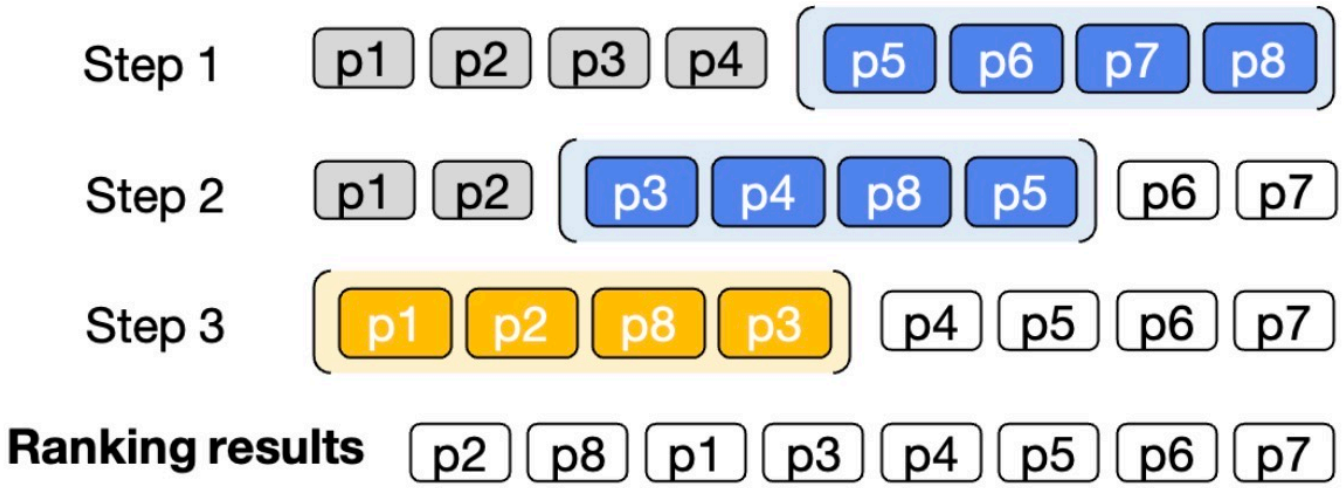
assistant:
Received passage [2]

(more passages) ...

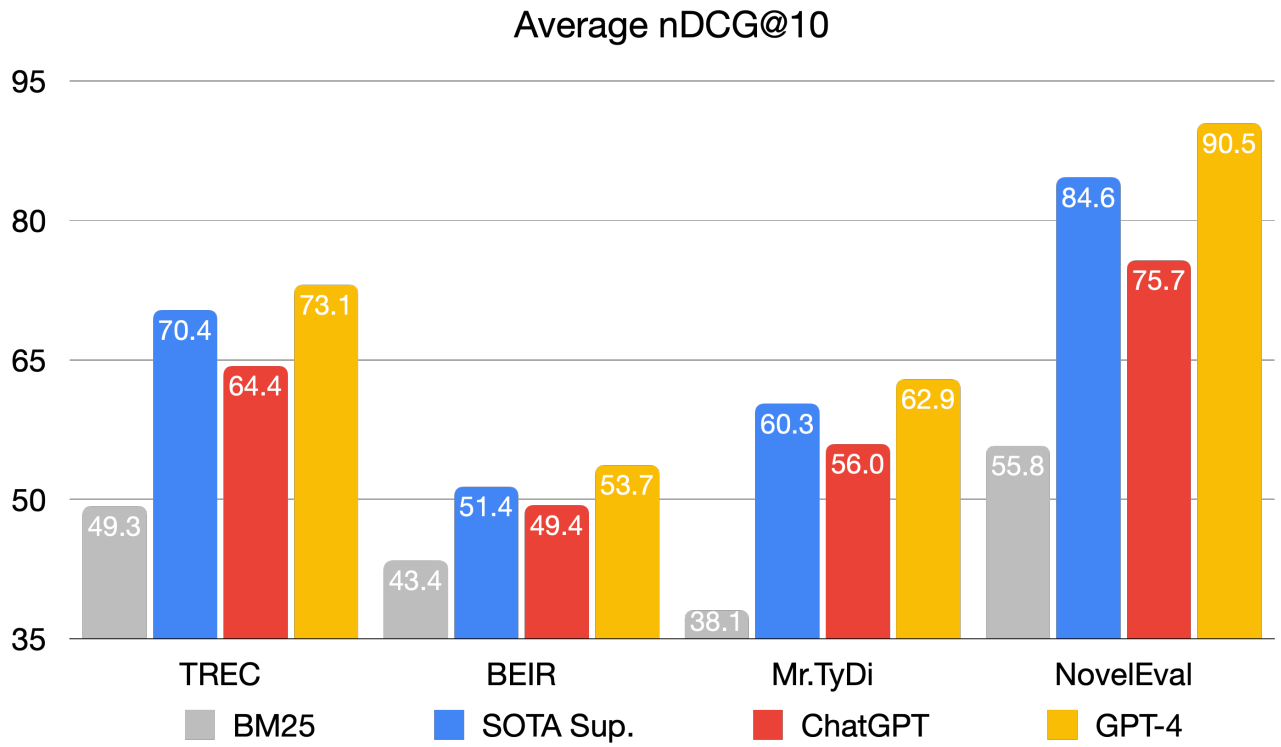
user
Search Query: {{query}}.
Rank the {{num}} passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers, and the most relevant passages should be listed first, and the output format should be [] > [], e.g., [1] > [2]. Only response the ranking results, do not say any word or explain.

Passage Ranking by LLMs

- LLM-based passage ranking
  - **Permutation Generation**
  - **Sliding Window Strategy**



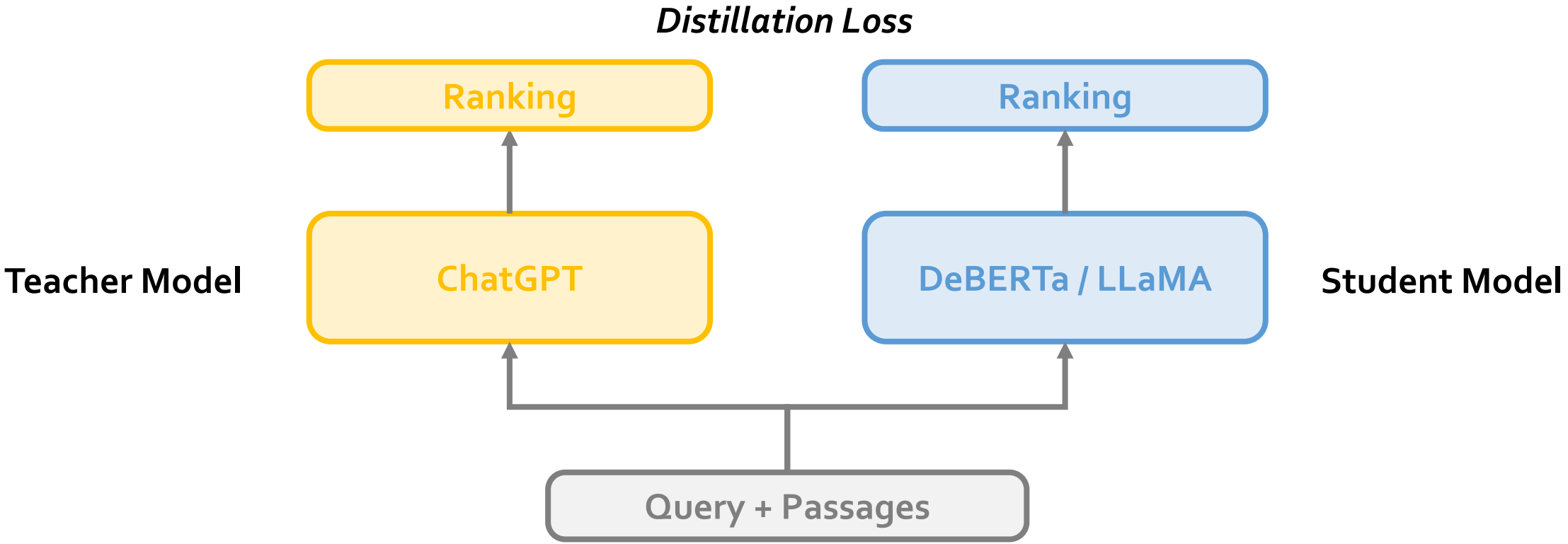
Passage Ranking by LLMs



Outperforms SOTA supervised models

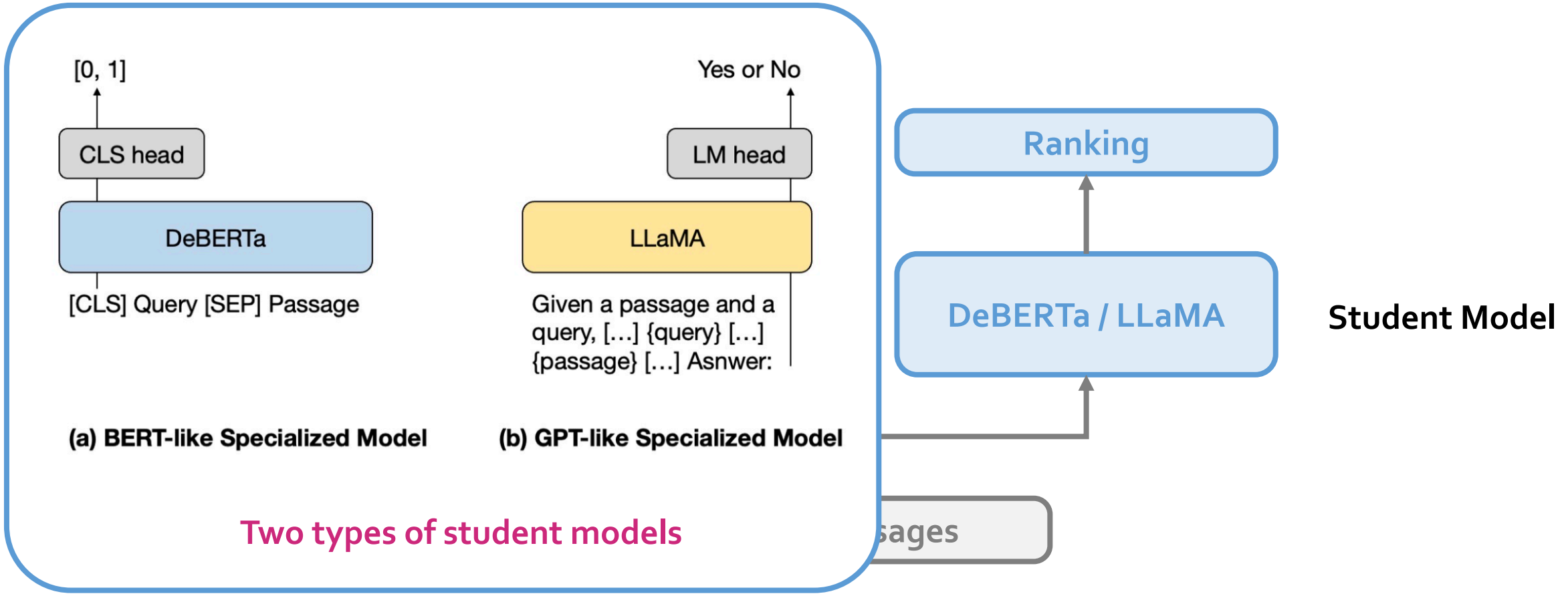
Passage Ranking by LLMs

- Permutation distillation



Passage Ranking by LLMs

- Permutation distillation



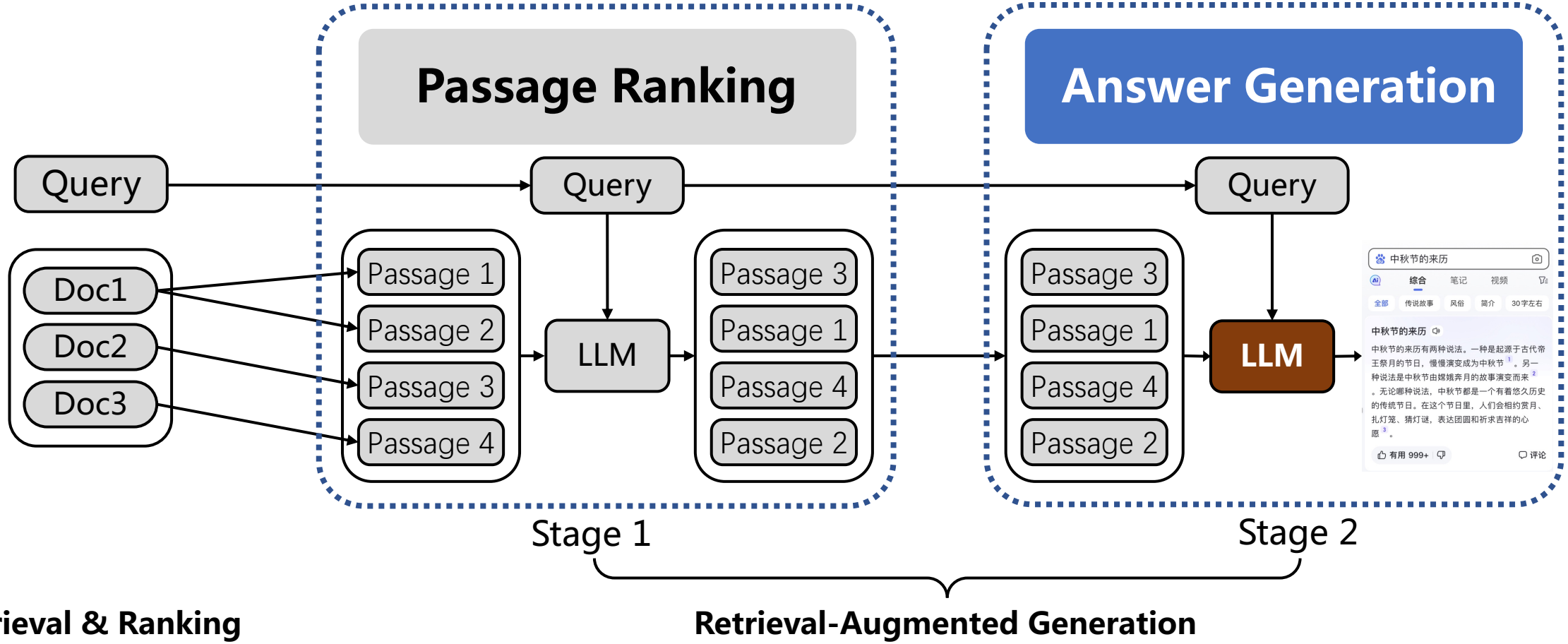
Passage Ranking by LLMs

- Evaluation Results (NDCG@10)

Label	Method	DL19	DL20	BEIR (Avg)
∅	BM25	50.58	47.96	43.42
∅	ChatGPT	65.80	62.91	49.37
MARCO	monoT5 (3B)	71.83	68.89	51.36
MARCO	DeBERTa-Large	68.89	61.38	42.64
MARCO	LLaMA-7B	69.24	58.97	47.71
ChatGPT	DeBERTa-Large	70.66	**67.15**	**53.03**
ChatGPT	LLaMA-7B	**71.78**	66.89	51.68

The distilled models outperforms ChatGPT and monoT5 on benchmarks

Retrieval-Augmented Generation in Baidu



Retrieval-Augmented Generation

- Main Challenges in RAG
  - Precisely understanding instruction for the complicate retrieval task

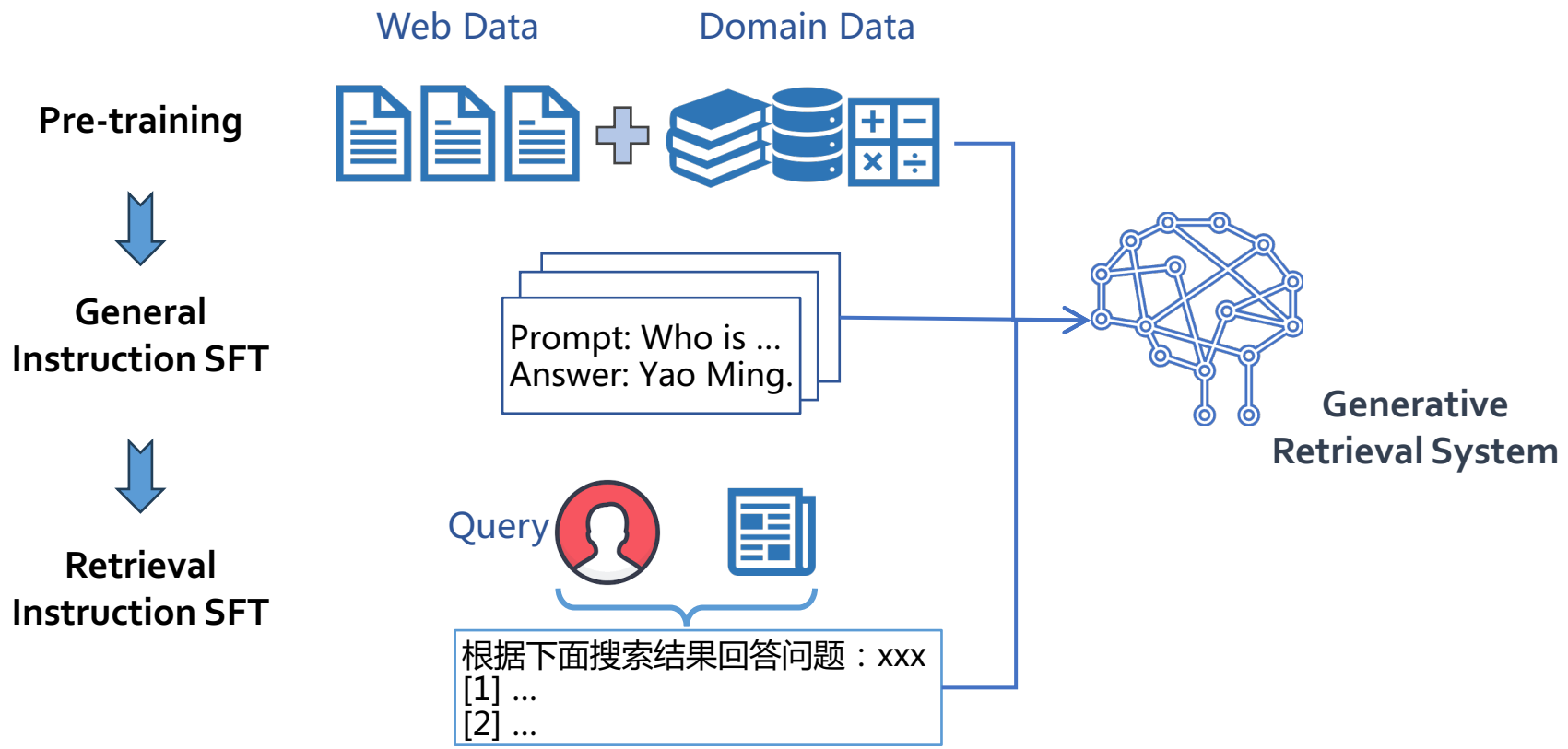
Retrieval-aware LLM training

  - Comprehensive generation combining LLM and retrieval knowledge

Learning from heterogeneous feedbacks

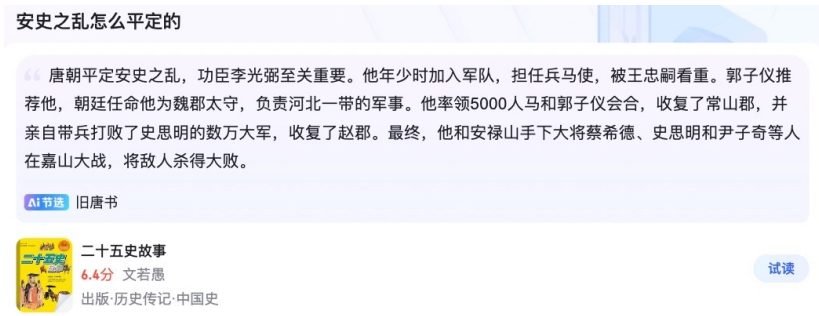
Retrieval-Augmented Generation in Baidu

- Retrieval-aware LLM training
  - Multi-stage

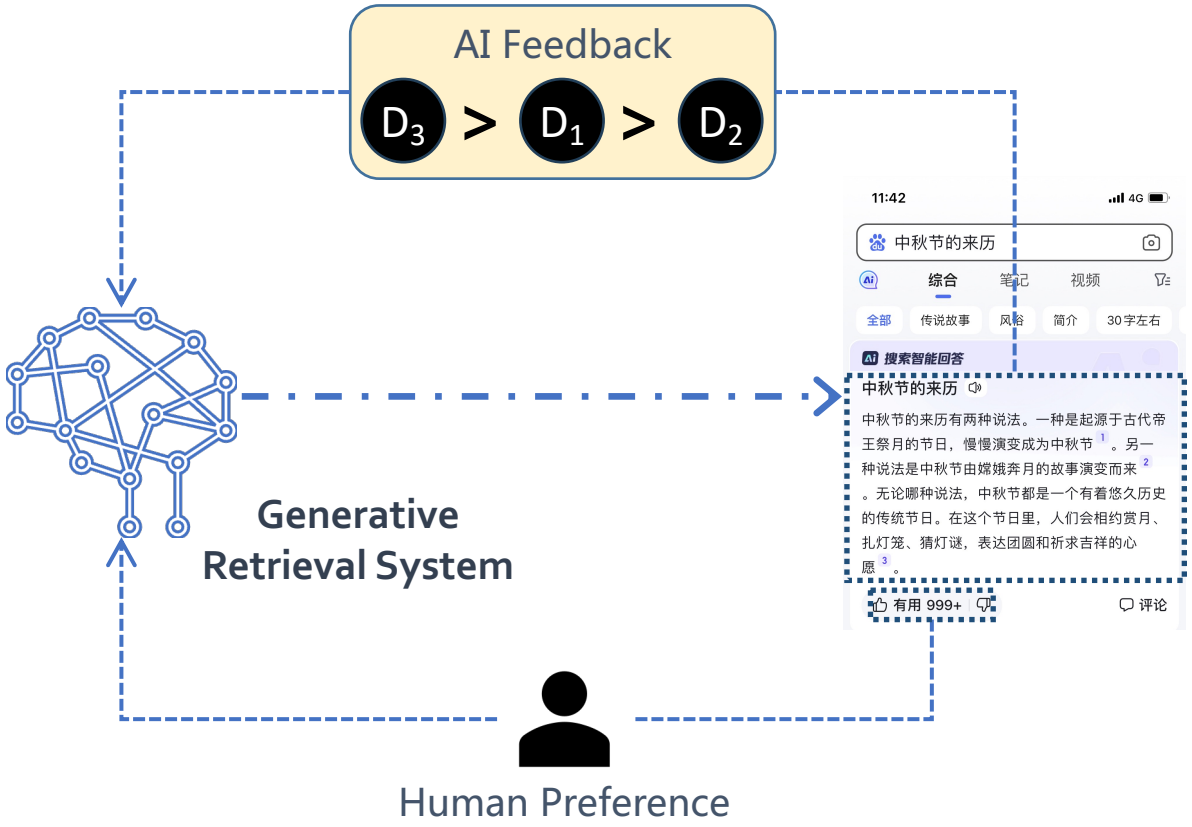


Retrieval-Augmented Generation in Baidu

- Retrieval-aware LLM training
  - Multi-source



Retrieval-Augmented Generation in Baidu



Future Work

Hallucinations in RAG:
- Hallucination detection[1]
- RLHF for hallucinations
- …

LLM Evaluation
- Evaluation Benchmarks[2]
- Evaluation Metrics
- …

Multi-modal RAG
- Multi-modal understanding
- Multi-modal generation
- …

[1] Zhao et al. Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. arXiv preprint *arXiv:2310.17918*.
[2] Zhao et al. DiQAD: A Benchmark Dataset for End-to-End Open-domain Dialogue Assessment. In *Findings of EMNLP 2023*.


OUR MISSION
MAKE THE COMPLICATED
WORLD SIMPLER THROUGH TECHNOLOGY
BAIDU
We are hiring!
Please drop a message to wangshuaiqiang@baidu.com if interested.
Baidu 百度