

# Retrieval-Augmented Generation via LLMs in Web Search

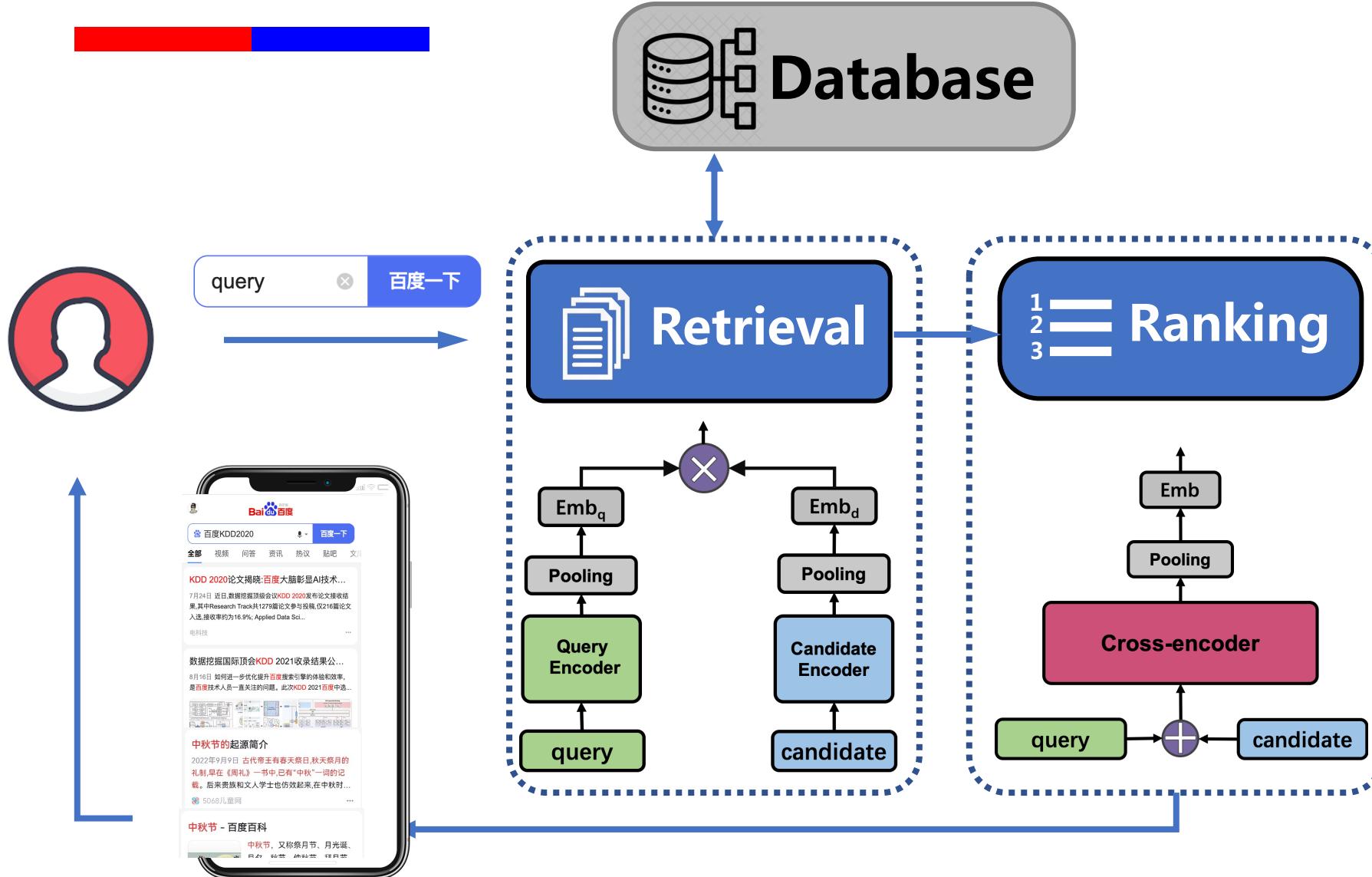
---

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



**09:07**

问一问

手机送话器声音小怎么办

摘要 您好, 如果您的手机送话器声音太小, 可以尝试以下几种方法来解决问题: 1. 调整手机音量: 在通话中, 您可以使用手机音量键来调整音量大小。请注意, 不同手机型号的音量键可能会有所不同, 您可以查看手机说明书或在手机设置中查找音量调节选项。2. 清洁送话器: 手机送话器可能会因为灰尘、污垢等原因而导致声音变小。您可以使用软刷或棉签轻轻擦拭手机送话器, 去除污垢和灰尘, 以提高声音的清晰度和响度。3. 关闭降噪功能: 在一些手机上, 有降噪功能会降低环境噪音, 但同时也会减小话筒的接收声音, 导致声音变小。可以尝试关闭降噪功能来解决问题。4. 检查软件设置: 在一些手机上, 可能存在软件设置问题, 导致送话器声音变小。您可以进入手机设置, 查看通话相关设置, 例如是否启用了“听筒模式”等, 以检查是否存在这些问题。如果以上方法都不能解决问题, 建议您联系手机厂商的客服部门或专业维修人员, 寻求进一步的帮助和支持。

咨询记录 · 回答于2023-04-25

手机送话器声音小怎么办

**09:07**

手机送话器声音小怎么办

AI 搜索智能回答

手机送话器声音小怎么办

1. 调整通话音量: 不同手机型号音量键不同, 可查看说明书或在手机设置中查找。
2. 清洁送话器: 用软刷或棉签轻轻擦拭, 提高清晰度和响度。
3. 关闭降噪功能: 部分手机有降噪功能会降低噪音, 可尝试关闭。
4. 检查软件设置: 通话相关设置中检查是否存在问题, 如“听筒模式”。如仍无解决, 建议联系客服或专业维修人员。

查看更多 >

有用 评论

大家还在搜

- 送话器声音小对方听... 怎么判断送话器坏...
- iphone送话器声音小... 手机送话器怎么清理...

**11:49**

陕西法制网 94.4万粉丝

中秋节来历

中秋节来历

中秋佳节由古代祭月慢慢演变而来。古时人们在八月十五这一天, 都会盼望她能相传, 就叫中秋, 所以

**11:39**

Q 中秋节的来历

中秋节由嫦娥奔月的故事演变而来。后羿射掉了多余的九个太阳, 然后上天赏赐他一种成仙的药, 他舍不得离开自己的妻子嫦娥, 就把药给了她。在八月十五这天, 人们都会盼望她能传下来, 所以叫中秋, 所以

**11:47**

Q 中秋节的来历

中秋佳节

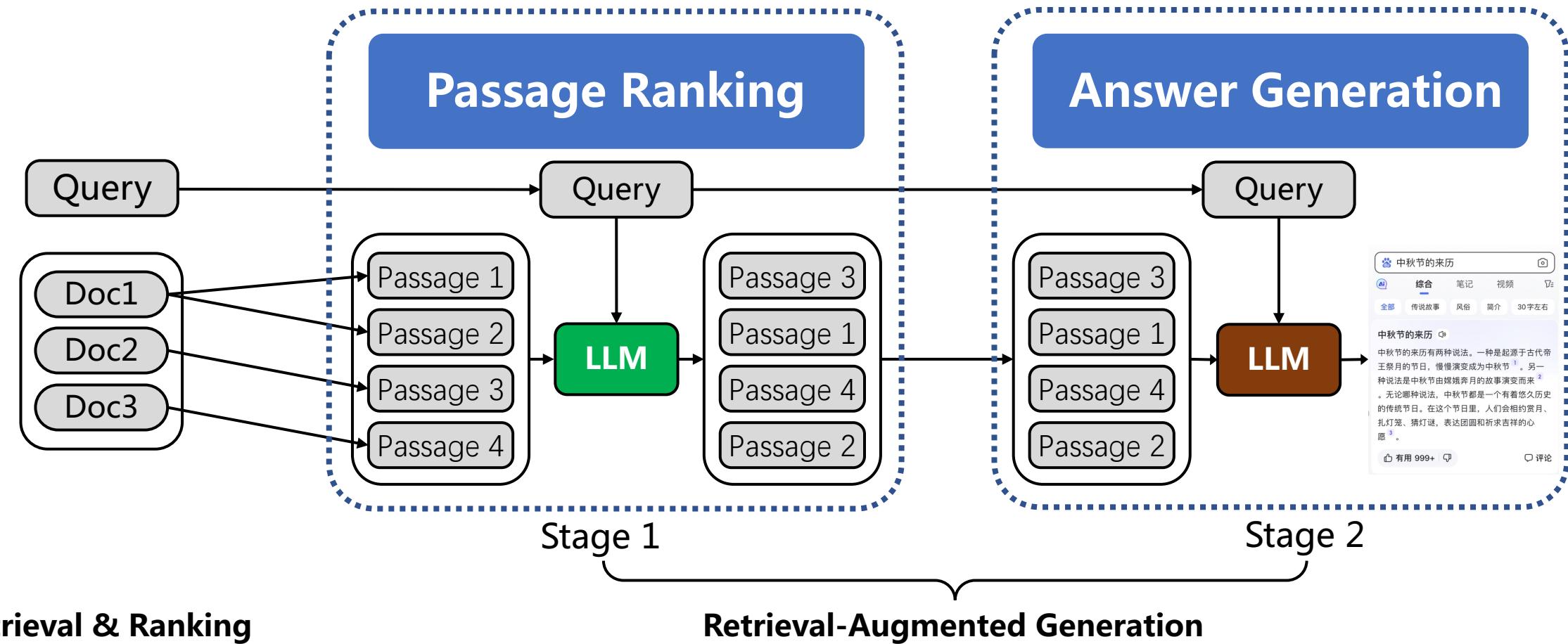
中秋佳节

中秋节起源于上古时代, 普及于汉代, 定型于唐朝初年, 盛行于宋朝以后。在这天, 朋友们相约赏月、扎灯笼、猜灯谜。中秋节到底是怎么来的?

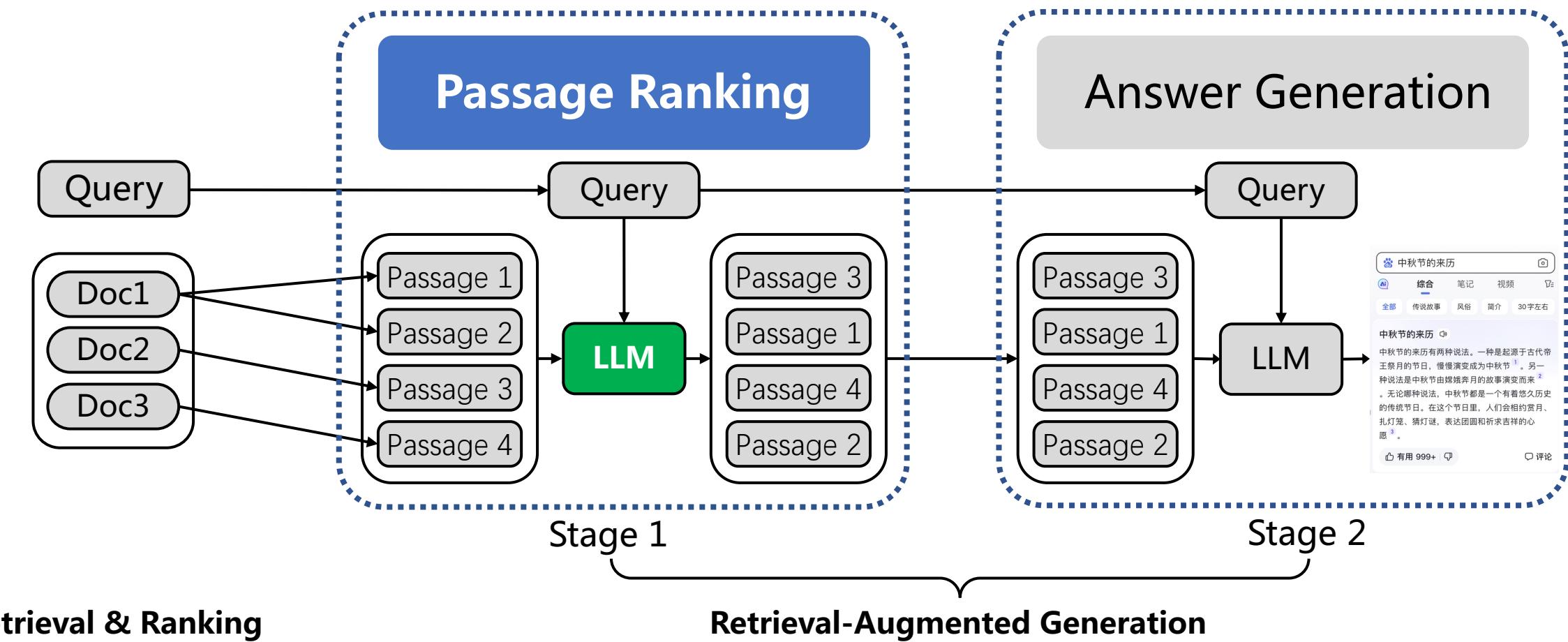
大家还在搜

- 中秋节的真正由来 中秋节的起源简短
- 中秋节的来历20个字 端午节的恐怖传说
- 中秋节的由来简介30字 中秋节的真正来历20字
- 中秋节三大风俗 中秋节来历简短40字

# Retrieval-Augmented Generation in Baidu

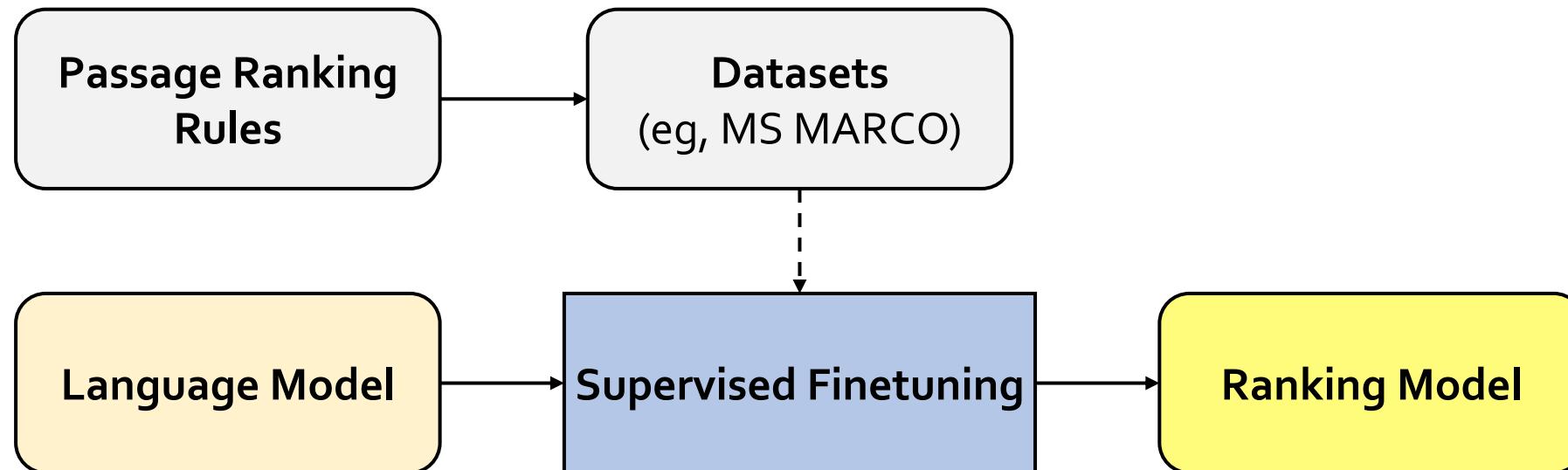


# Retrieval-Augmented Generation in Baidu



# Passage Ranking

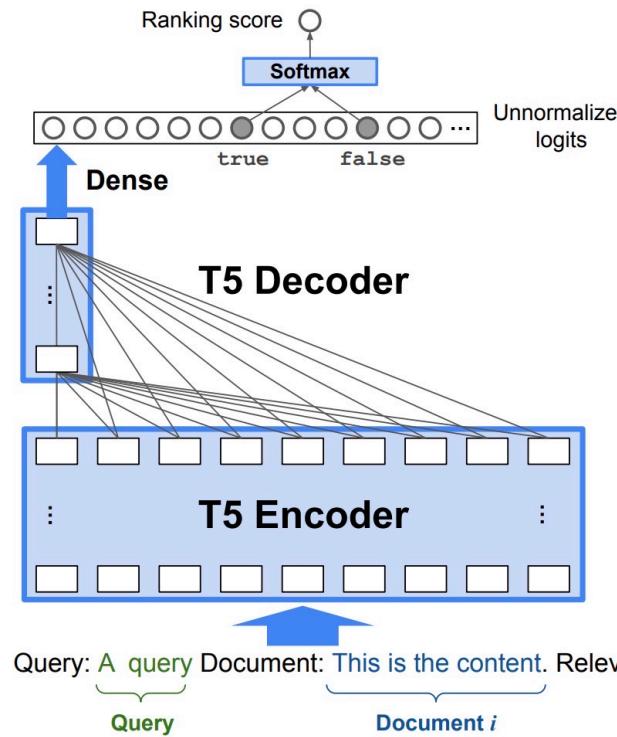
- Fine-tuning based passage ranking



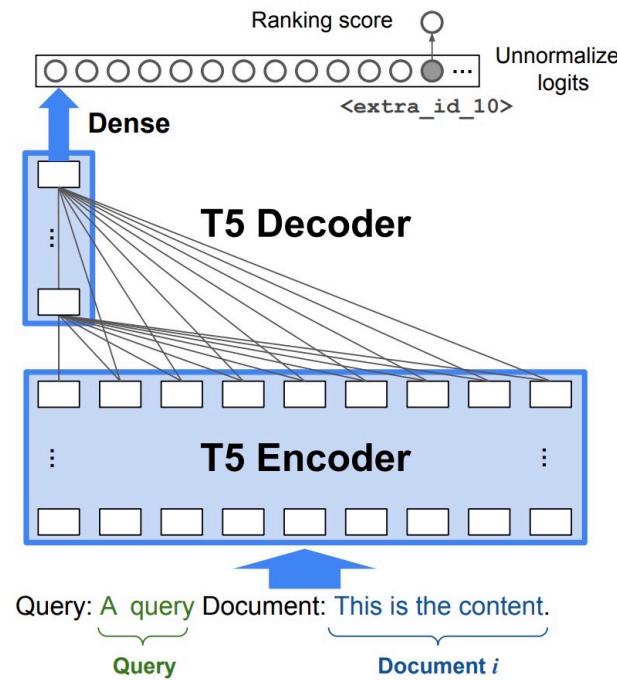
- [1] Nogueira et al. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *EMNLP (Findings) 2020*.
- [2] Zhuang et al. RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses. In *SIGIR 2023*.

# Passage Ranking by LLMs

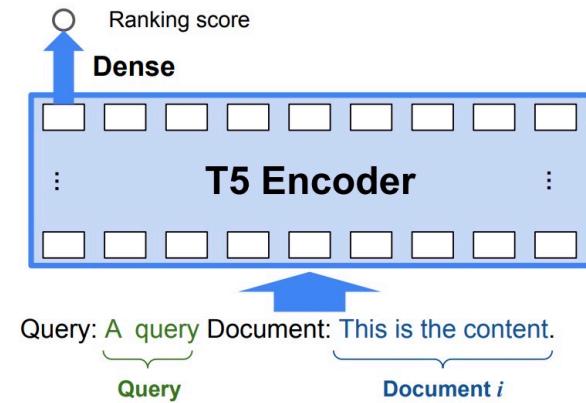
- Fine-tuning based ranking



(a) monoT5 (Nogueira et al., 2020)



(b) RankT5 (Encoder-Decoder)



(c) RankT5 (Encoder-Only)

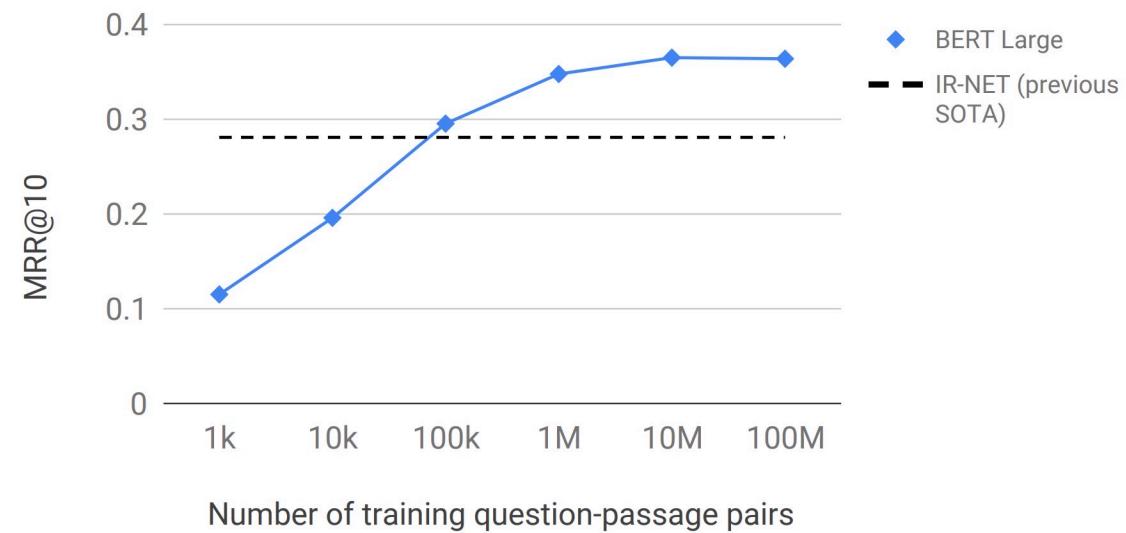
- [1] Nogueira et al. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *EMNLP (Findings) 2020*.  
[2] Zhuang et al. RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses. In *SIGIR 2023*.

# Passage Ranking by LLMs



- Challenges of fine-tuning based ranking

Significant human effort

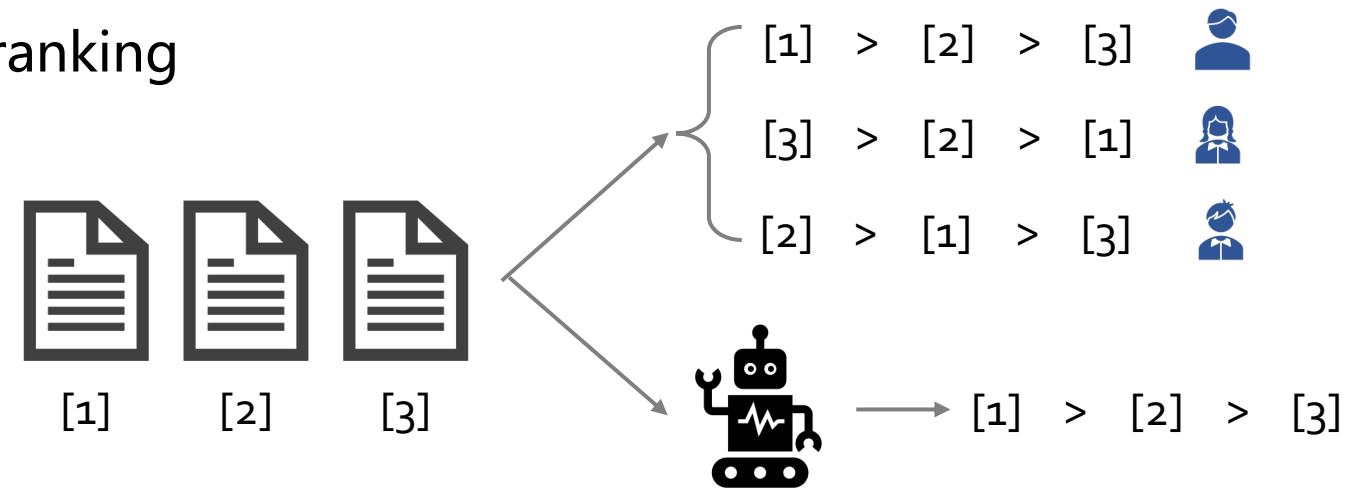


# Passage Ranking by LLMs

- Challenges of fine-tuning based ranking

Significant human effort

Inconsistent annotation



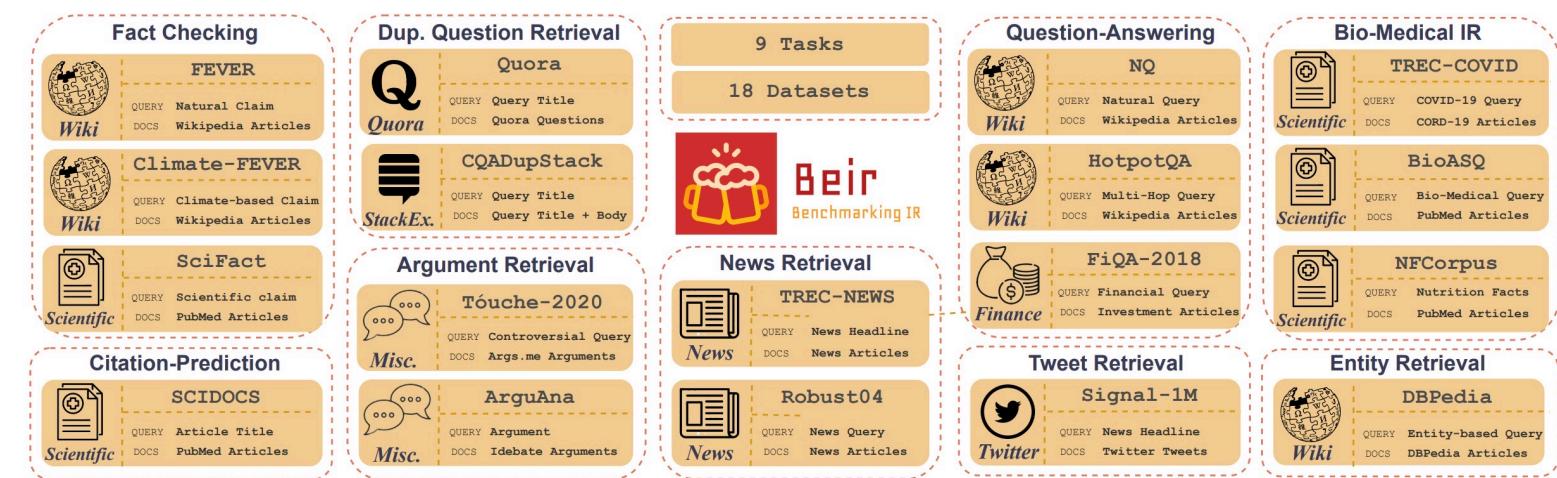
# Passage Ranking by LLMs

- Challenges of fine-tuning based ranking

Significant human effort

Inconsistent annotation

Weak generalizability



Diverse Ranking Tasks

# Passage Ranking by LLMs

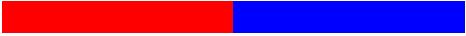


- LLM-based passage ranking



**Ranking with prompt by LLMs**

# 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) ...

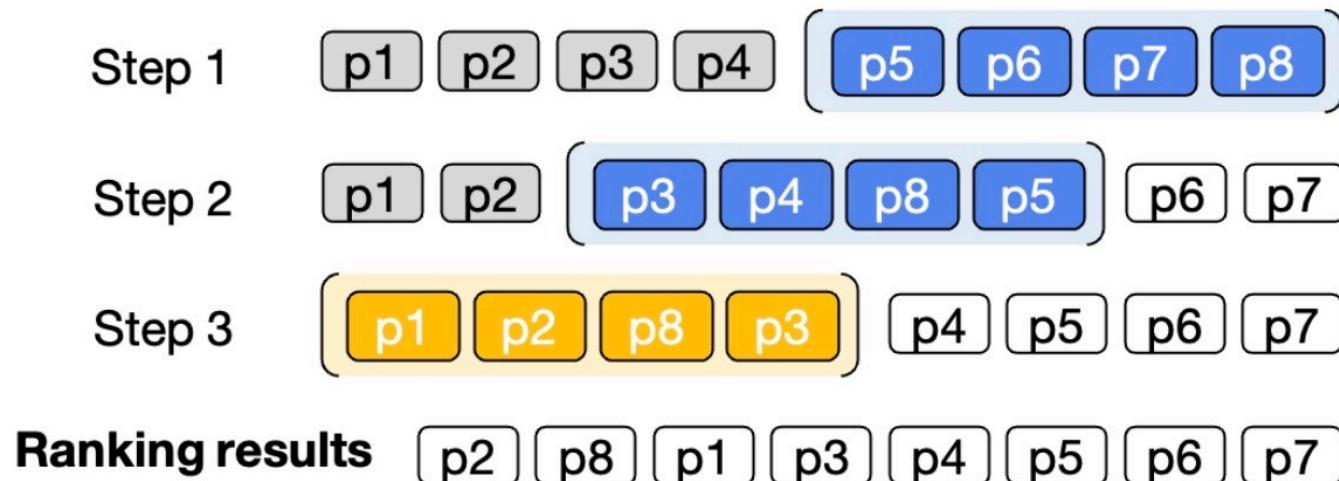
**user**

Search Query: {{query}}.

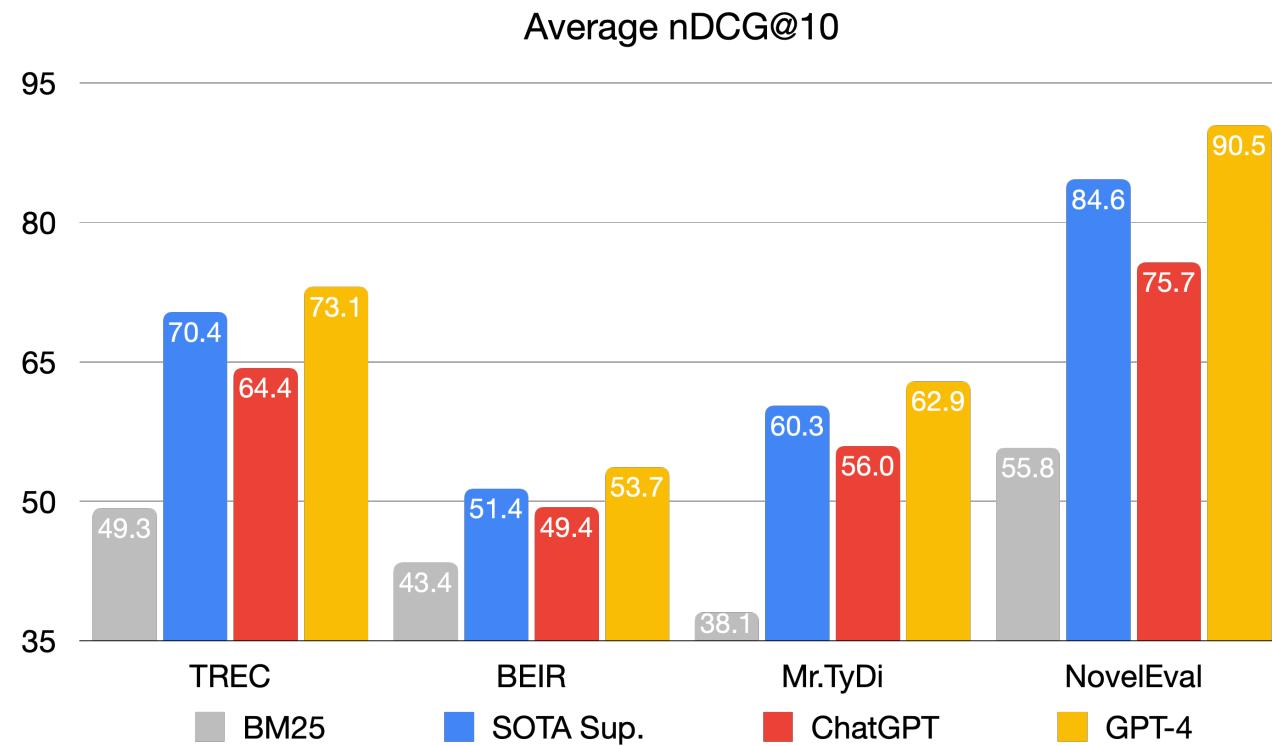
Rank the {{num}} passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers, and the most relevant passages should be listed first, and the output format should be [] > [], e.g., [1] > [2]. Only response the ranking results, do not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

# Passage Ranking by LLMs

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



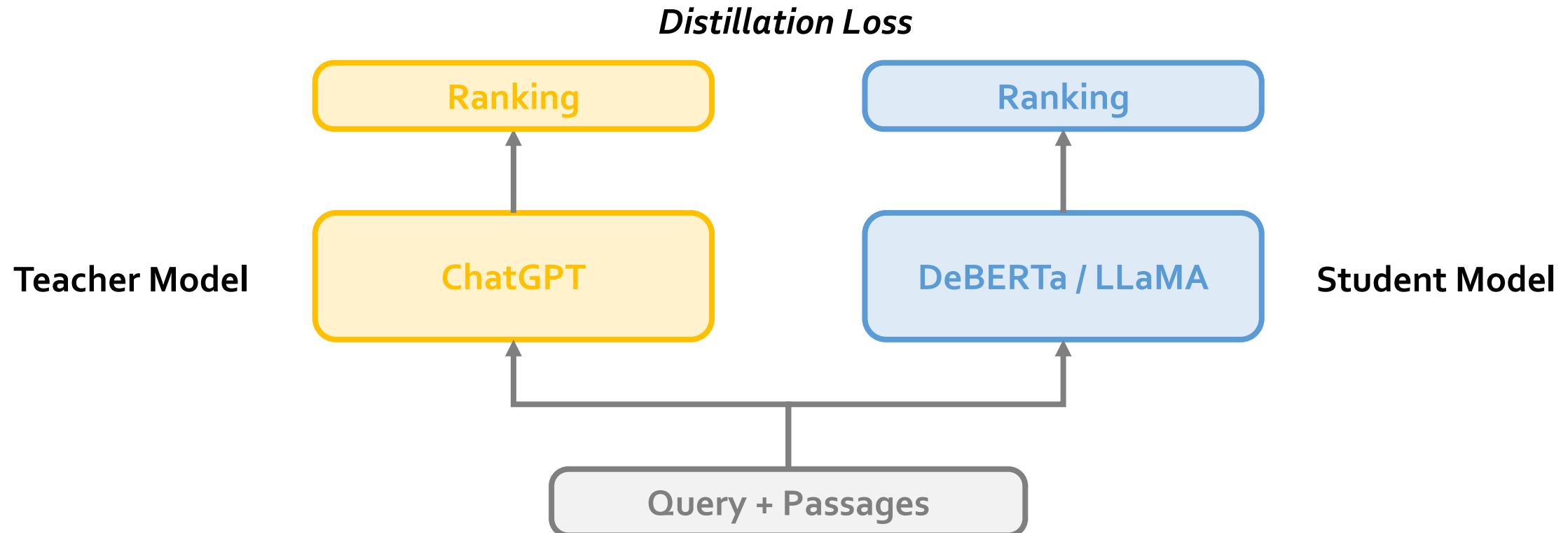
# Passage Ranking by LLMs



Outperforms SOTA supervised models

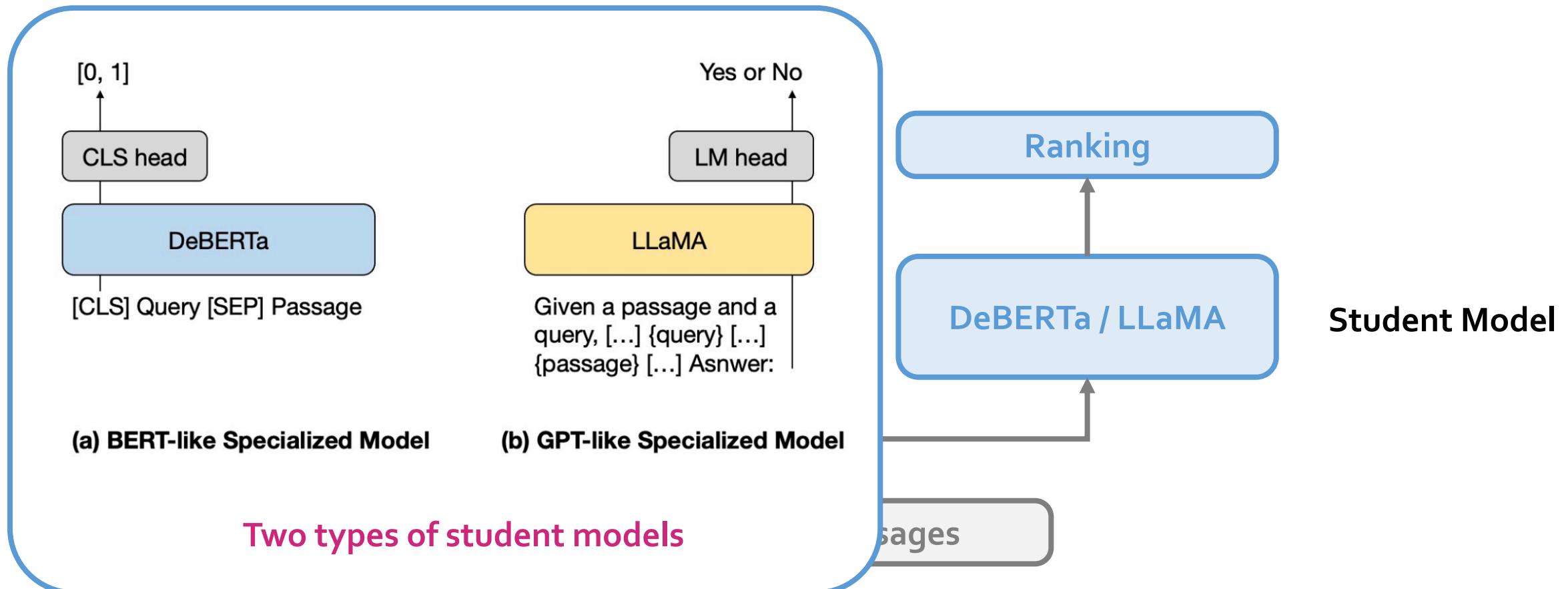
# Passage Ranking by LLMs

- Permutation distillation



# Passage Ranking by LLMs

- Permutation distillation



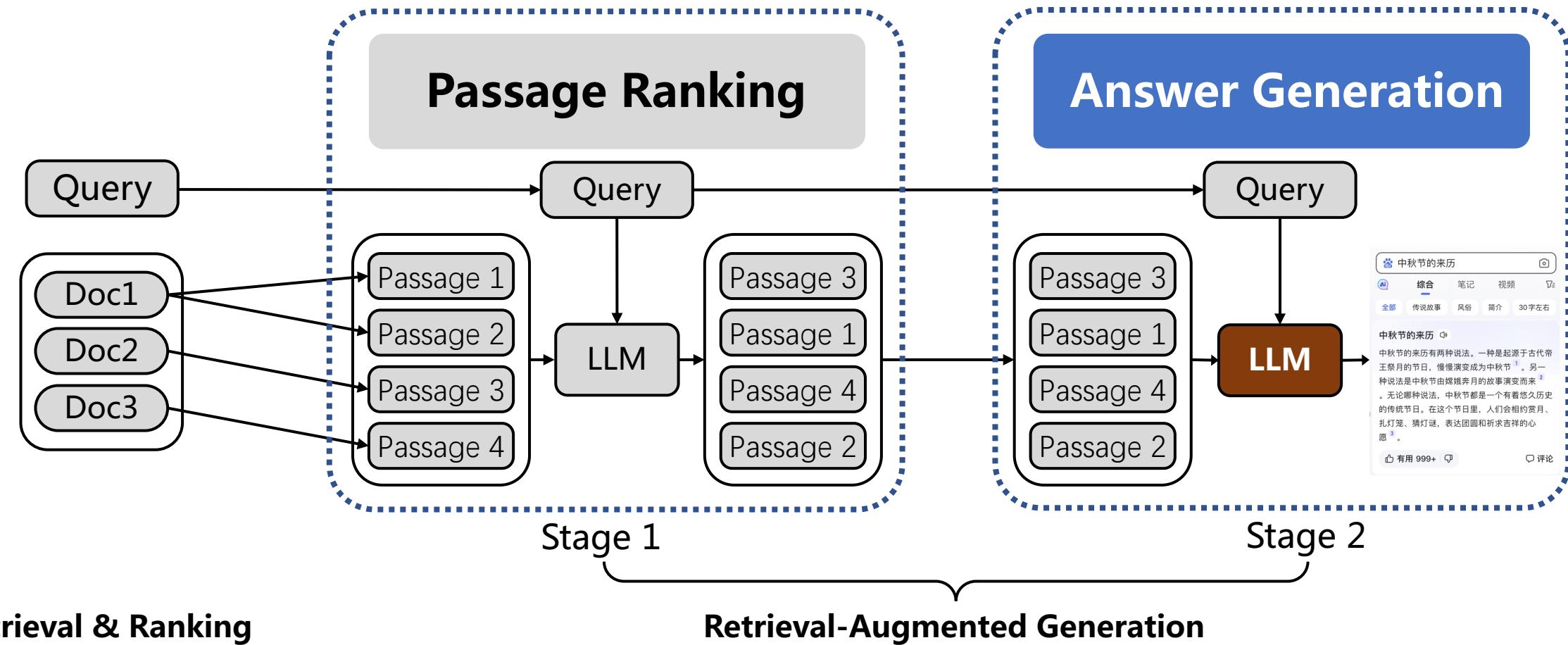
# Passage Ranking by LLMs

- Evaluation Results ( NDCG@10 )

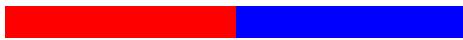
Label	Method	DL19	DL20	BEIR (Avg)
∅	BM25	50.58	47.96	43.42
∅	ChatGPT	65.80	62.91	49.37
MARCO	monoT5 (3B)	71.83	68.89	51.36
MARCO	DeBERTa-Large	68.89	61.38	42.64
MARCO	LLaMA-7B	69.24	58.97	47.71
ChatGPT	DeBERTa-Large	70.66	<b>67.15</b>	<b>53.03</b>
ChatGPT	LLaMA-7B	<b>71.78</b>	66.89	51.68

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

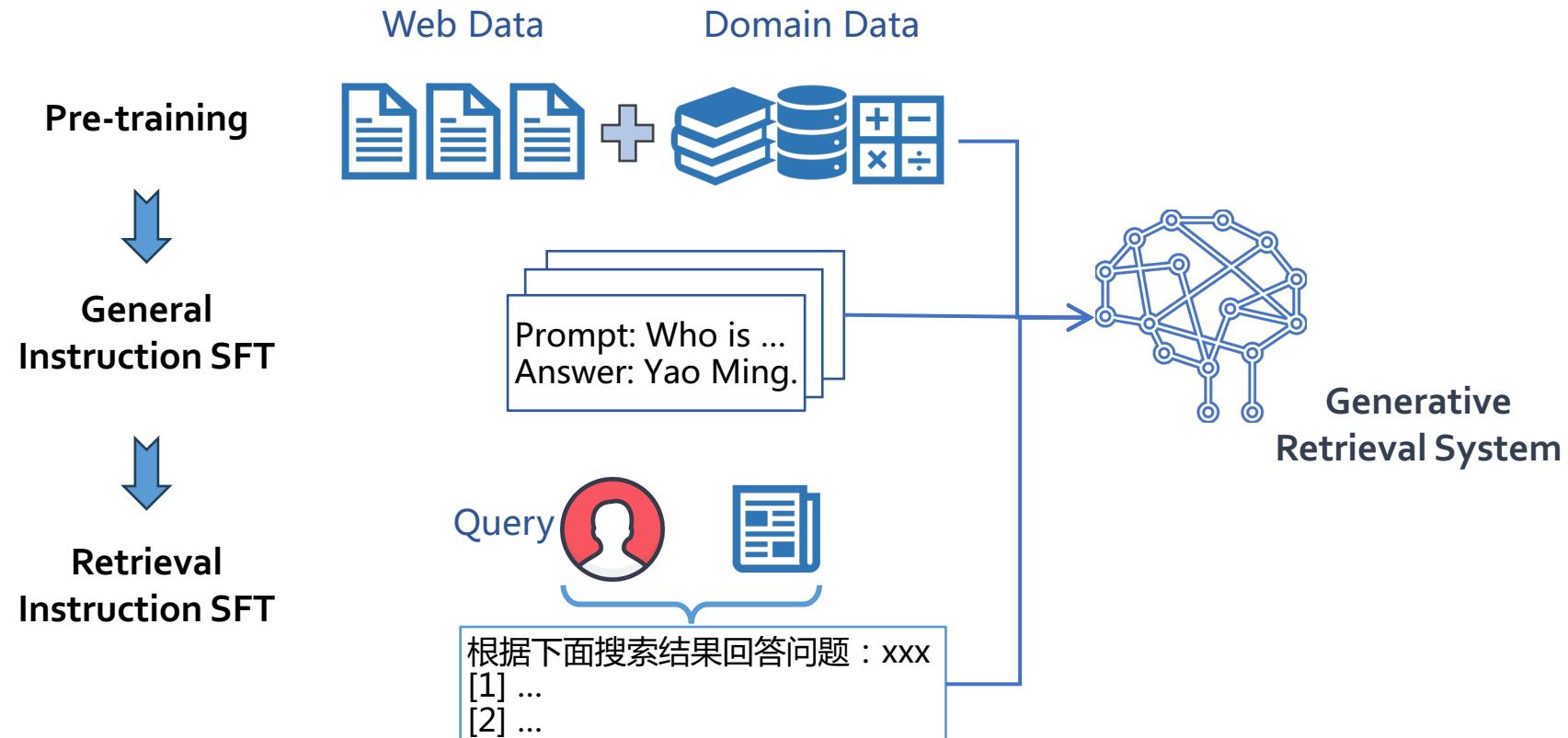
## **Retrieval-aware LLM training**

- 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

**中秋节的起源和由来**



2022年12月2日 据史籍记载,古代帝王祭月的节期为农历八月十五,时日恰逢三秋之半,故名“中秋节...”

应届毕业生网

**安史之乱怎么平定的**



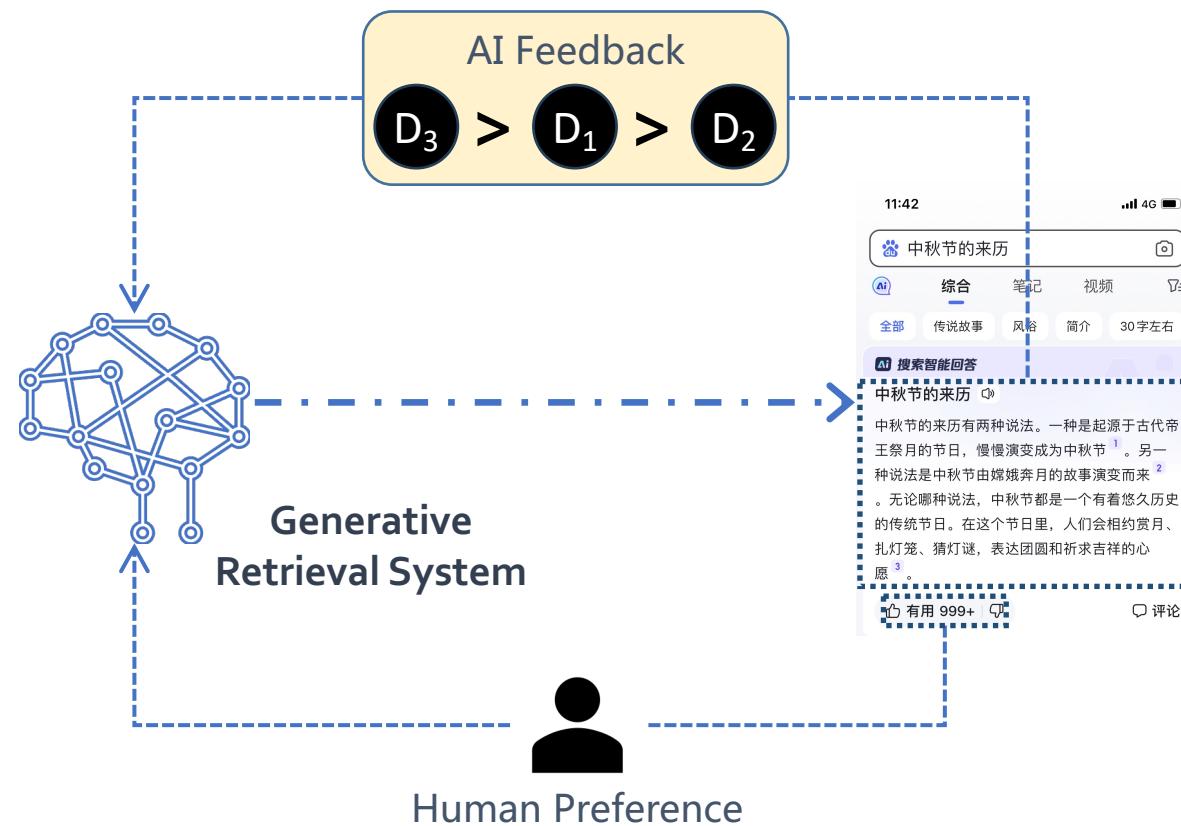
唐朝平定安史之乱,功臣李光弼至关重要。他年少时加入军队,担任兵马使,被王忠嗣看重。郭子仪推荐他,朝廷任命他为魏郡太守,负责河北一带的军事。他率领5000人马和郭子仪会合,收复了常山郡,并亲自带兵打败了史思明的数万大军,收复了赵郡。最终,他和安禄山手下大将蔡希德、史思明和尹子奇等人在嘉山大战,将敌人杀得大败。

AI节选 旧唐书 试读

地域类别	地域	时间类别	时间	人口数	单位
地级	唐山市	年度	2021年	769.7	万人
直辖市	北京市	年度	2022年	2184.3	万人
地级	大同市	年度	2021年	309.9	万人

章节 人口数 SEP 动物学院 爱他就了解他 地域类别：地级 地域：石家庄市 时间类别：

# Retrieval-Augmented Generation in Baidu



# Future Work



## Hallucinations in RAG:

- Hallucination detection<sup>[1]</sup>
- RLHF for hallucinations
- ...

## LLM Evaluation

- Evaluation Benchmarks<sup>[2]</sup>
- 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



*We are hiring!*

Please drop a message to [wangshuaiqiang@baidu.com](mailto:wangshuaiqiang@baidu.com) if interested.