AI Box

检索增强生成

《大语言模型》编写团队：李军毅

检索增强生成

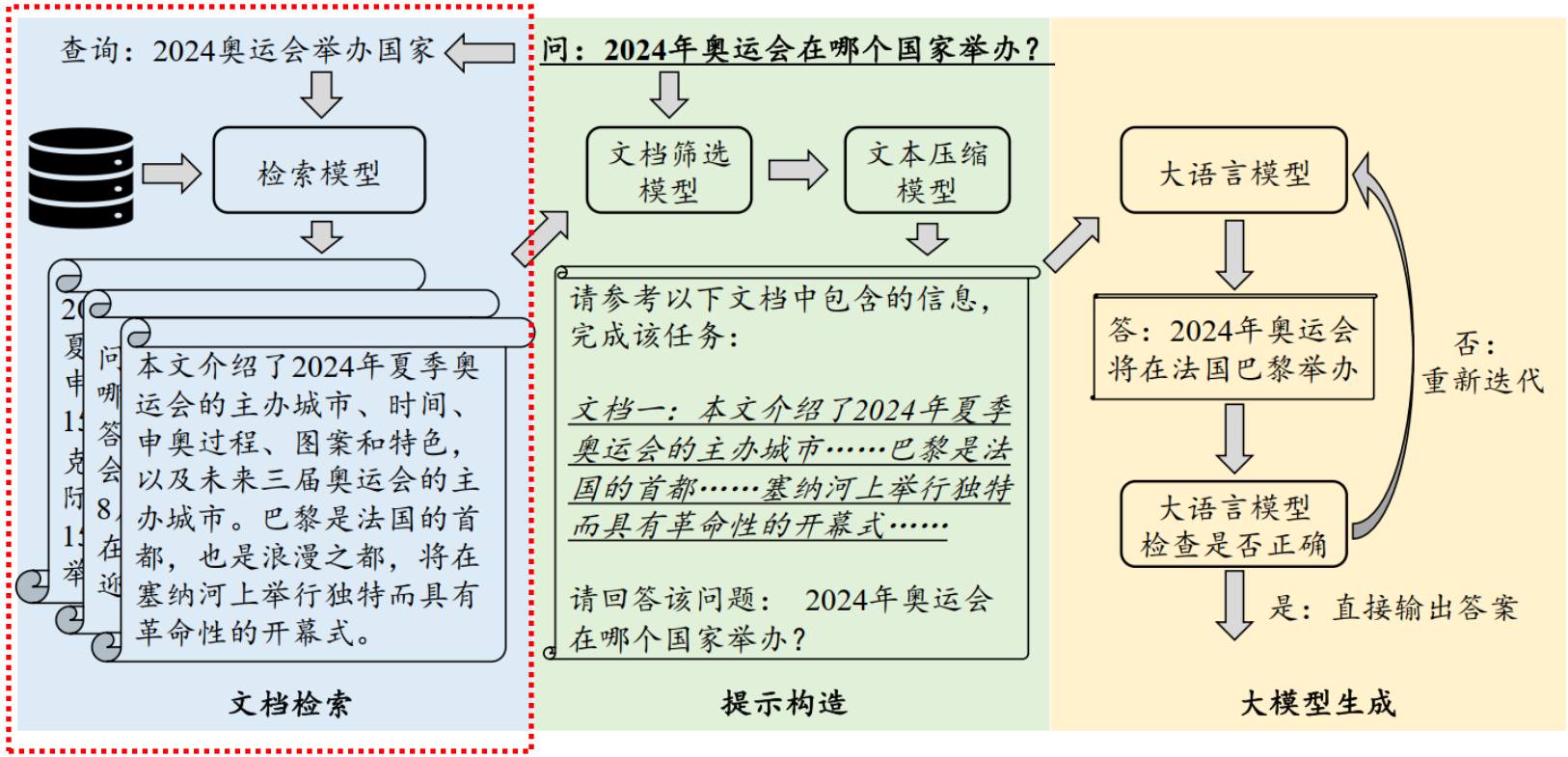
AI Box

- 大模型局限性
  - 幻觉
  - 时效性
  - 缺乏专业领域深度知识
  - 参数化知识效率低
- 实际应用的需求
  - 领域精准问答
  - 数据更新频繁
  - 生成内容可溯源

基本流程

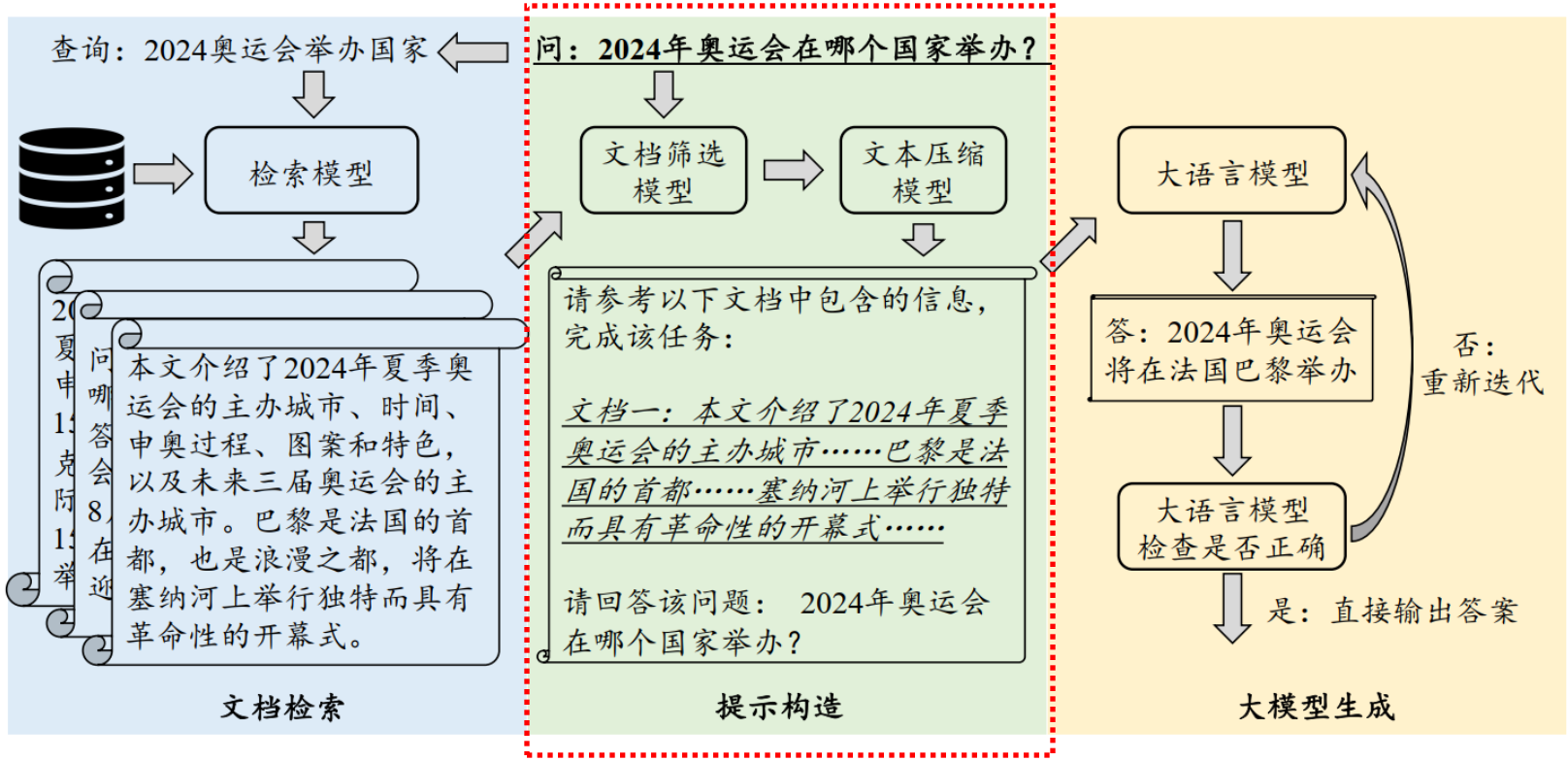
AI Box

- 三个阶段
  - 文档检索：从已有信息源中检索与问题相关的文档



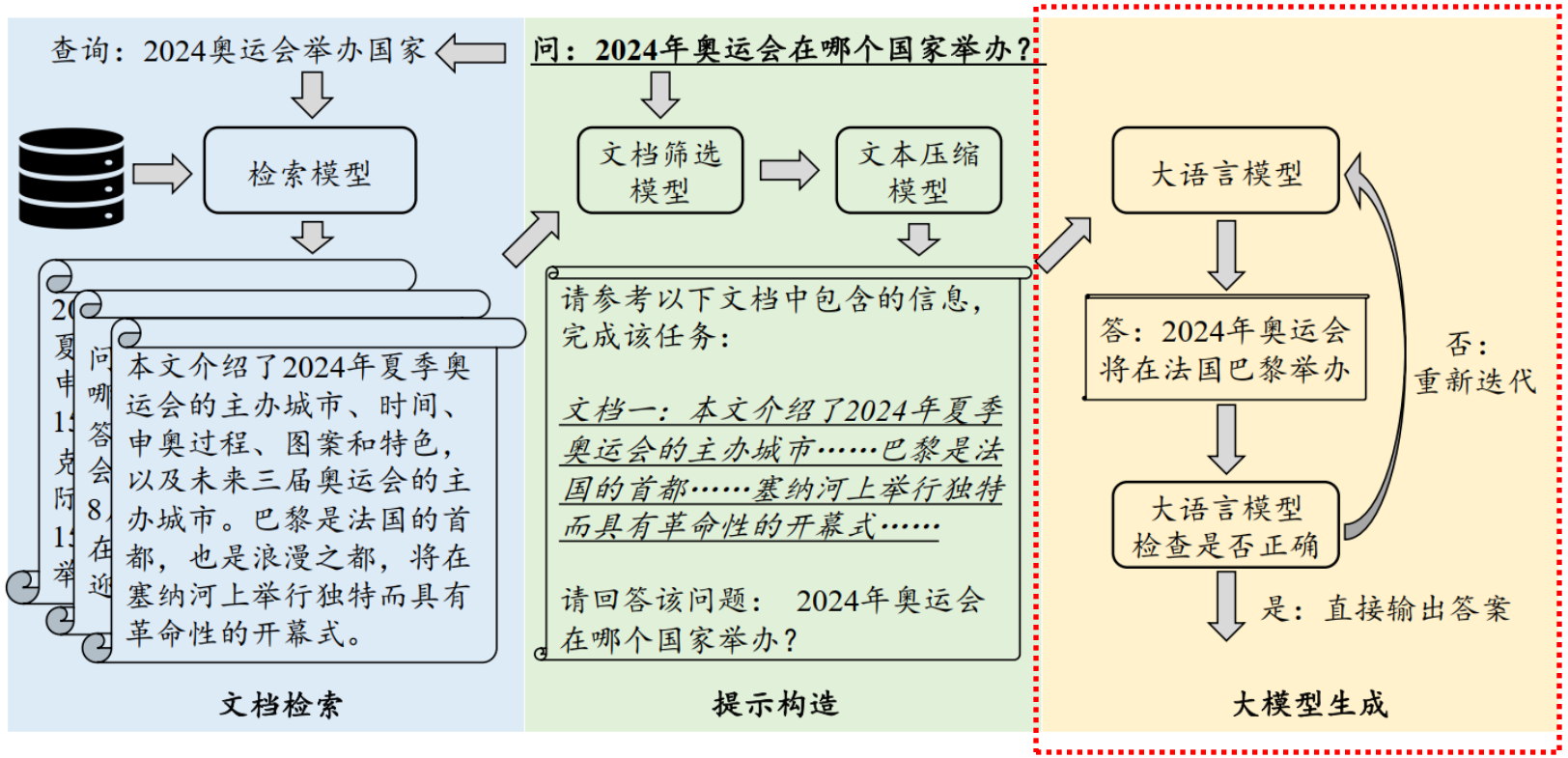
基本流程

- 三个阶段
  - 提示构造：将检索文档加入到输入提示中，并添加对应的任务描述



基本流程

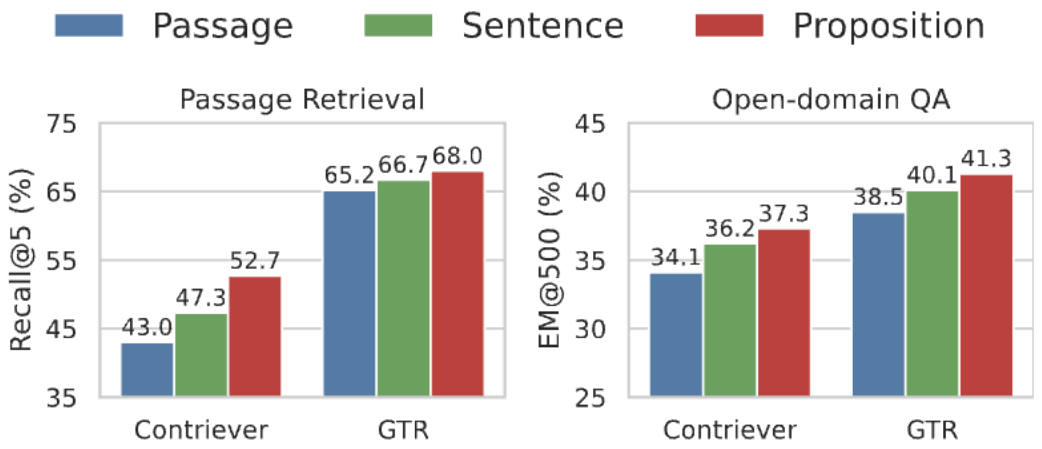
- 三个阶段
  - 大模型生成：模型基于检索增强的提示生成回答



优化策略

- 检索优化策略
  - 检索数据源优化
    - 检索单元：观点(语义完整且内容相对独立的文本片段)，需要训练专有模型提取观点
    - 对检索内容重组与改写，突出重要信息

文档级检索包含较多的无关信息



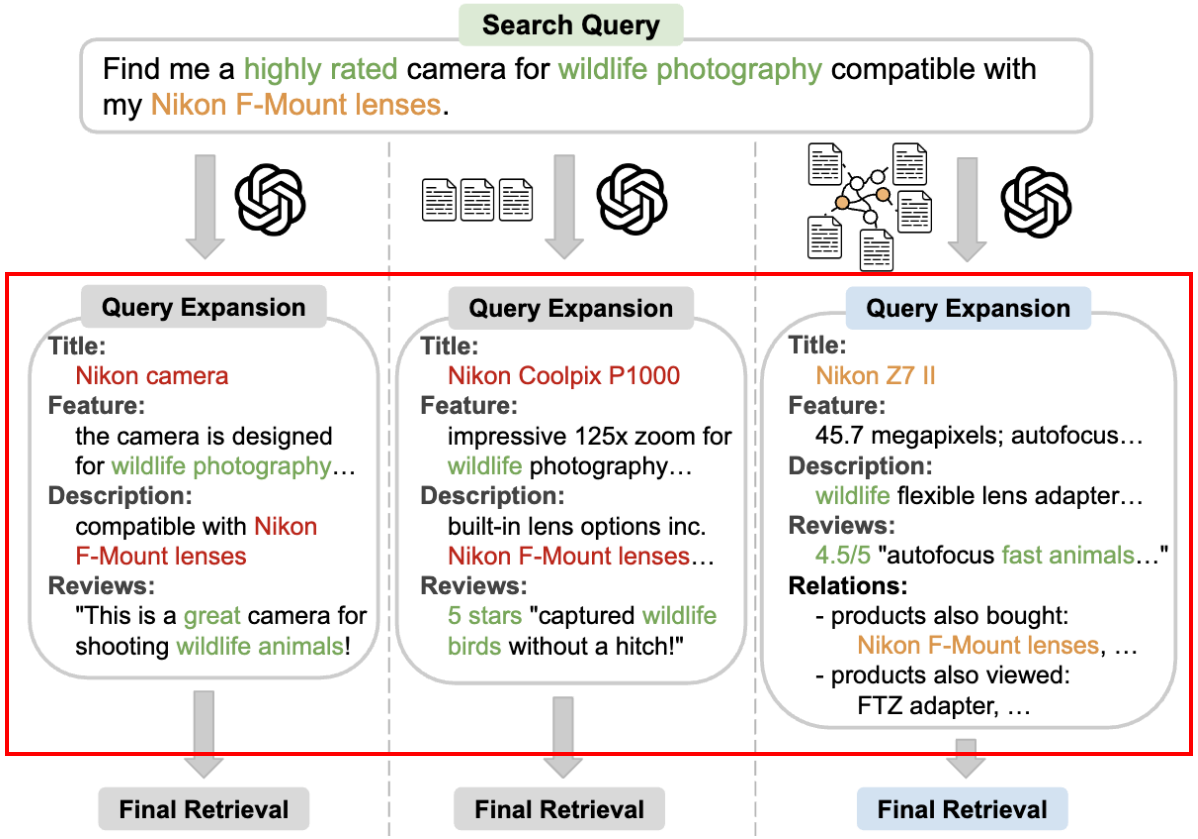
句子级检索召回更多相关信息，但检索时延较高

Dense X Retrieval: What Retrieval Granularity Should We Use?

优化策略

AI Box

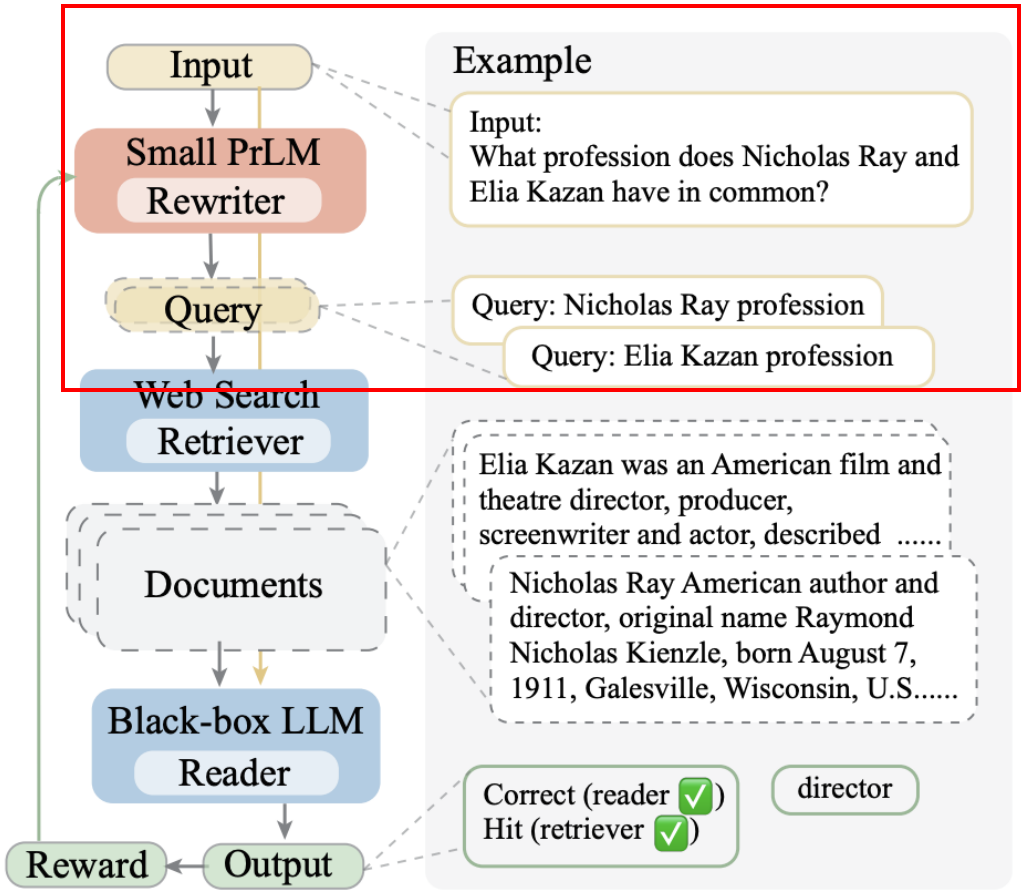
- 检索优化策略
  - 查询优化
    - 查询扩展
      - 为原始查询添加补充信息，对于复杂查询则可以分解为若干子查询
      - 每个子查询分别进行检索，并将中间结果合并



优化策略

AI Box

- 检索优化策略
  - 查询优化
    - 查询重写
      - 使关键信息更为突出，并消除可能存在的歧义
      - 可以训练专门的小模型用于查询重写，减少计算开销

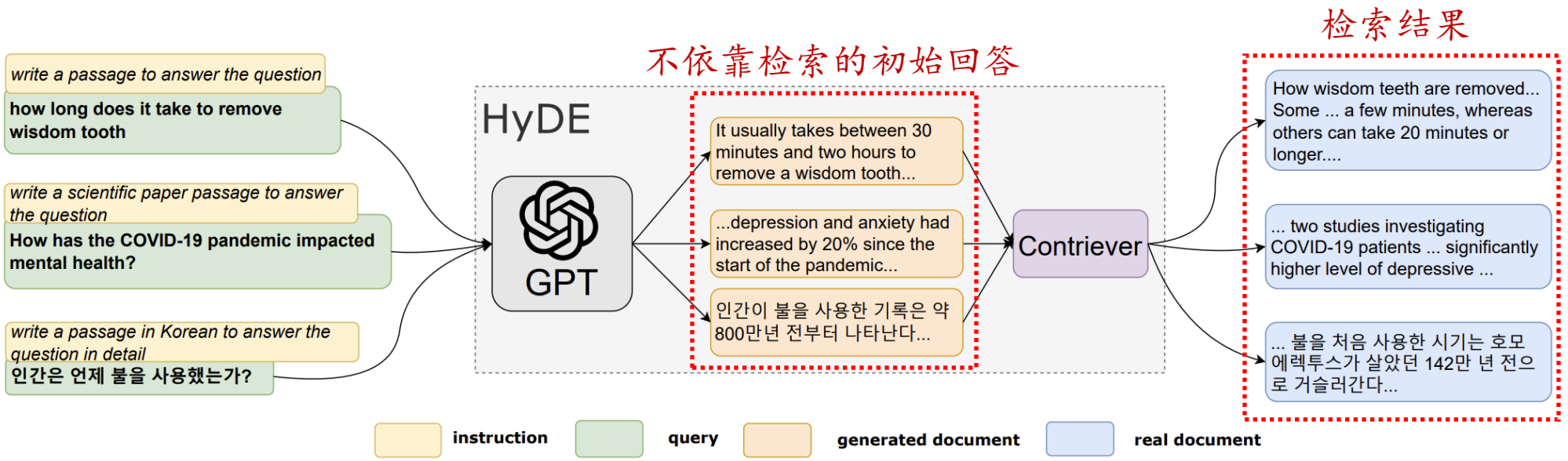


Query Rewriting for Retrieval-Augmented Large Language Models. **EMNLP 2023**

优化策略

AI Box

- 检索优化策略
  - 查询优化
    - 假设查询：用初始回答作为伪查询，检索得到新内容重新生成

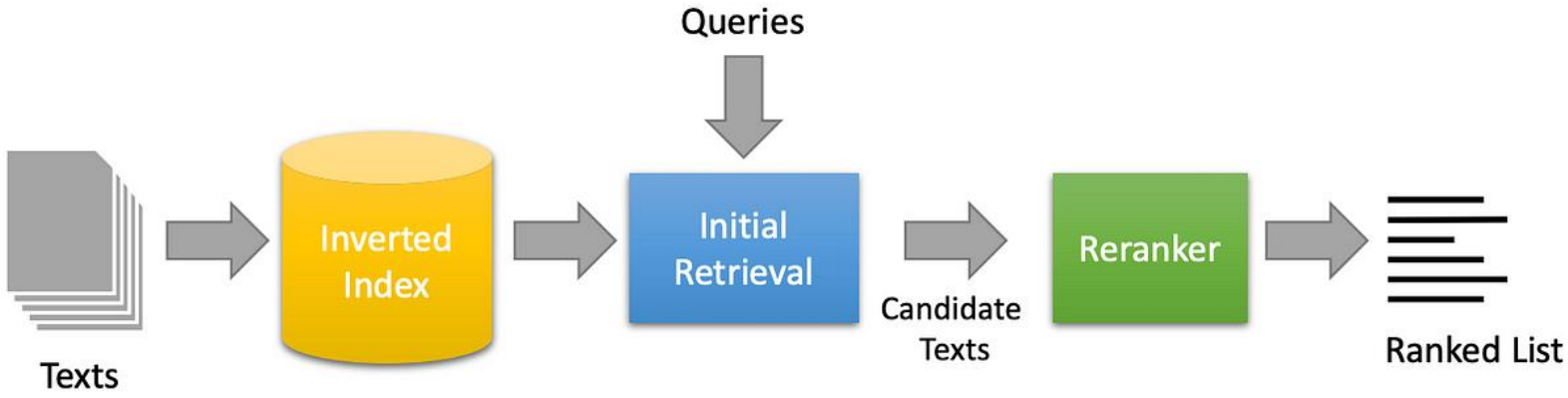


Precise Zero-Shot Dense Retrieval without Relevance Labels. **ACL 2023**

优化策略

AI Box

- 提示优化策略
  - 文档重排序
    - 根据文档与输入间的相关程度进行重新排序，过滤掉低质量或不相关的文档，或将相关度较低的文档放置在提示中的非最优位置 (如中间和靠后位置)
    - 可将大模型利用检索文档的情况转换为反馈分数用于训练重排序模型，提升重排序模型感知大模型对检索文档的利用能力



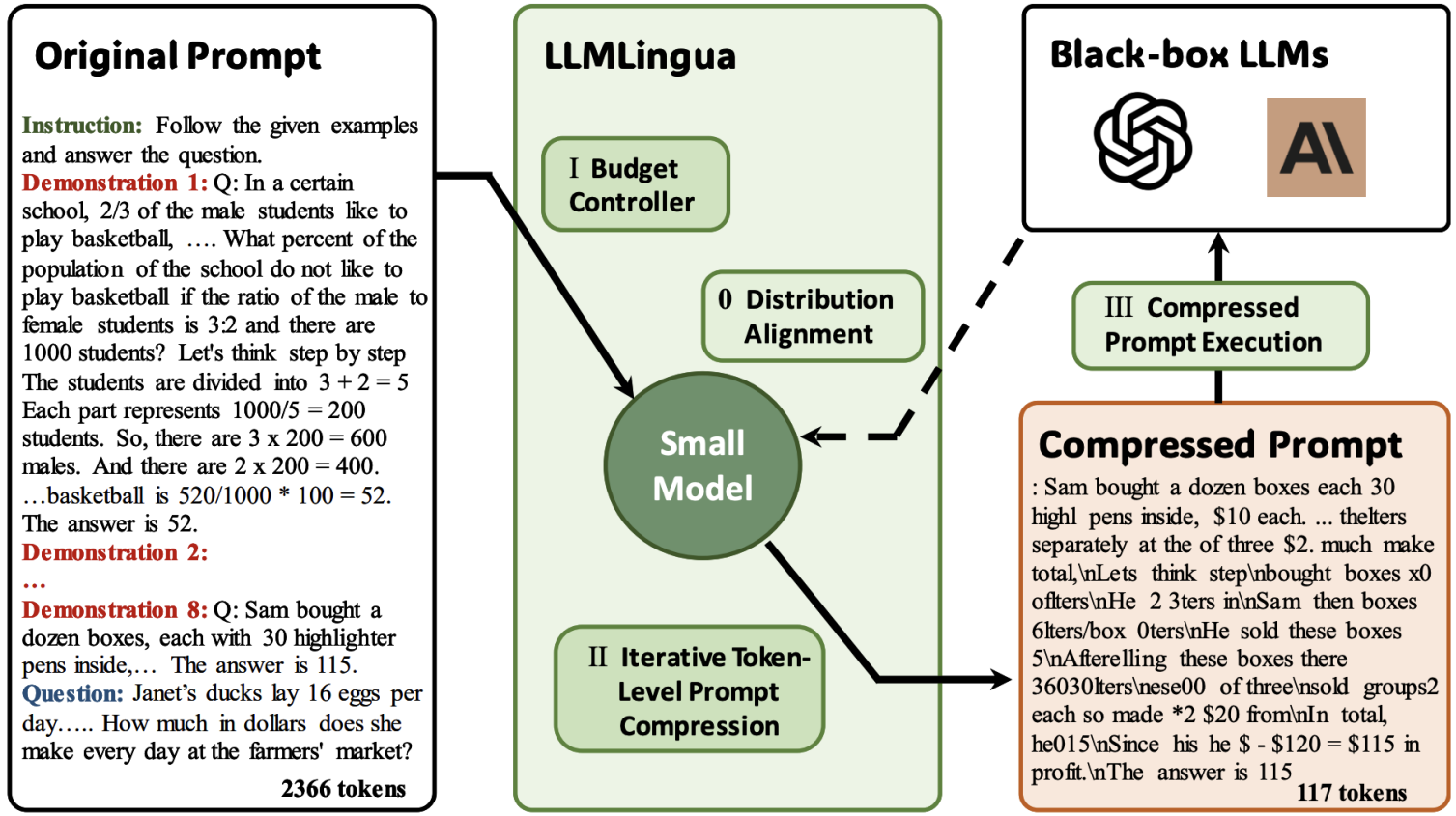
Document Ranking using Bert

优化策略

AI Box

- 提示优化策略
  - 上下文压缩
    - 自动摘要或信息抽取技术
    - 词元级别的压缩策略

根据大模型对词元的预测概率选择出重要的词元

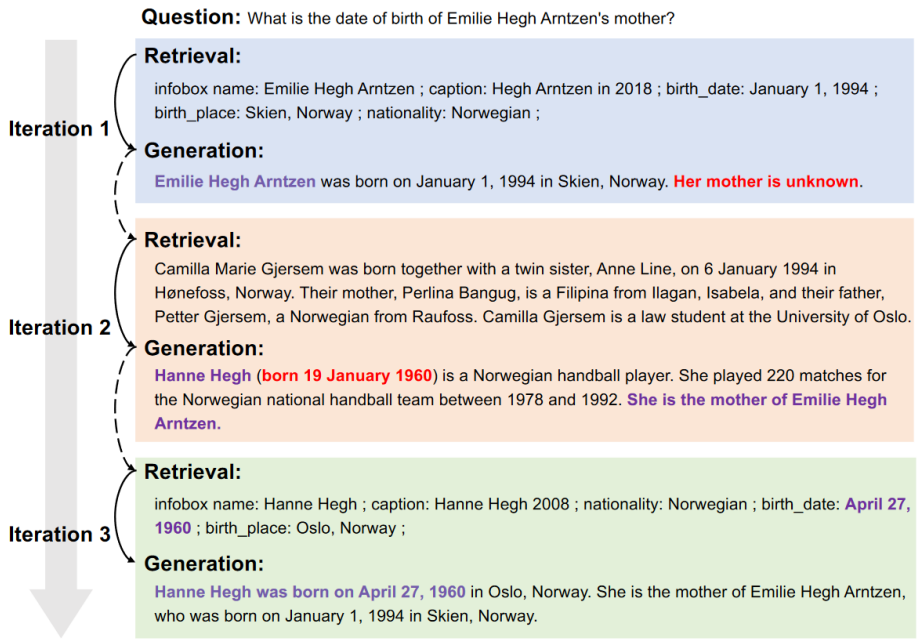


LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models

优化策略

AI Box

- 过程优化策略
  - 迭代检索增强：迭代合并模型输出与初始查询，多次检索
    - 检索也可与思维链结合，将思考步骤作为查询进行多步检索



Identified domains: factual (Wikidata, Wikipedia)

Rationale 1: First, the Argentine actor who directed El Tio Disparate is Fernando Birri.
Retrieve (Wikidata) 1: SELECT ?answer WHERE { wd:El Tio Disparate wdt:director ?answer . } -> Palito Ortega
Retrieve (Wikipedia) 1: Who directed El Tio Disparate? ->
El Tio Disparate is directed by Palito Ortega.
Corrected rationale 1: the Argentine actor who directed El Tio Disparate is Palito Ortega.

Rationale 2: Second, Palito Ortega was born in 1941.
Retrieve (Wikidata) 2: SELECT ?answer WHERE { wd:Palito Ortega wdt:date of birth ?answer . } -> 8 March 1941
Retrieve (Wikipedia) 2: When was Palito Ortega born? -> Palito Ortega was born in 8 Match 1941.
Corrected rationale 2: Palito Ortega was born in 8 Match 1941.

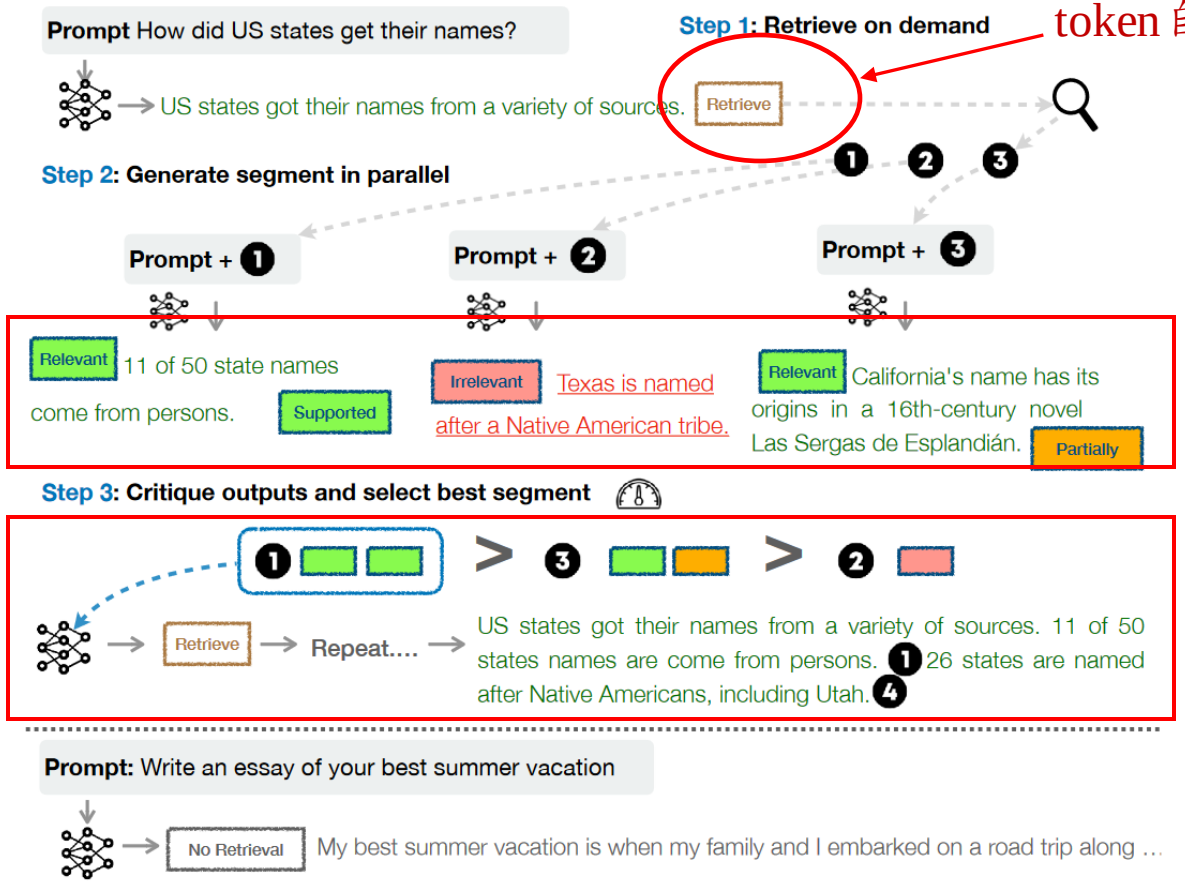
Corrected rationales: First, the Argentine actor who directed El Tio Disparate is Palito Ortega. Second, Palito Ortega was born in 8 Match 1941.

The answer is 1941.

Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources

优化策略

- 过程优化策略
  - 自适应检索增强
    - 大语言模型自主判断：
      - 何时需要使用检索器
      - 如何进行查询
      - 如何处理检索结果



大模型输出Retrieve token自主决定检索

大模型判断检索结果的相关性

大模型自主决定使用哪些文档

Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection, ICLR 2024

优化策略

AI Box

- 其他优化策略
  - 指令微调增强策略
    - 位置去偏 (相关文档置于不同位置)
    - 无关信息过滤 (添加无用文档)
  - 预训练增强策略
    - 针对性加强检索生成能力

将文档首段作为查询，训练模型根据检索结果生成文档剩余内容

(b) Search-Augmented Prompt

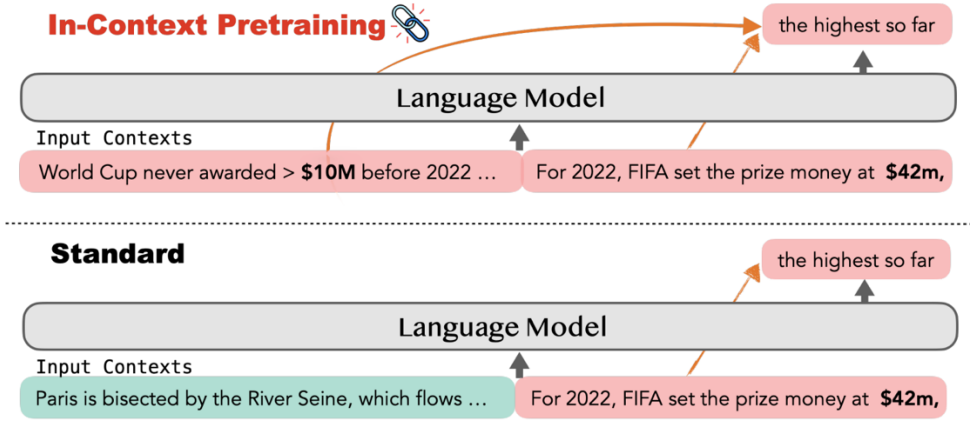
```
Below is an instruction that describes a task.
Write a response that appropriately completes the request.

### Related Information:
[Title 3]\n [Preview 3]
[Title 2]\n [Preview 2]
[Title 1]\n [Preview 1]

### Instruction: [Instruction]
### Input: [Input or None]

### Response:
```

在指令中加入外部信息

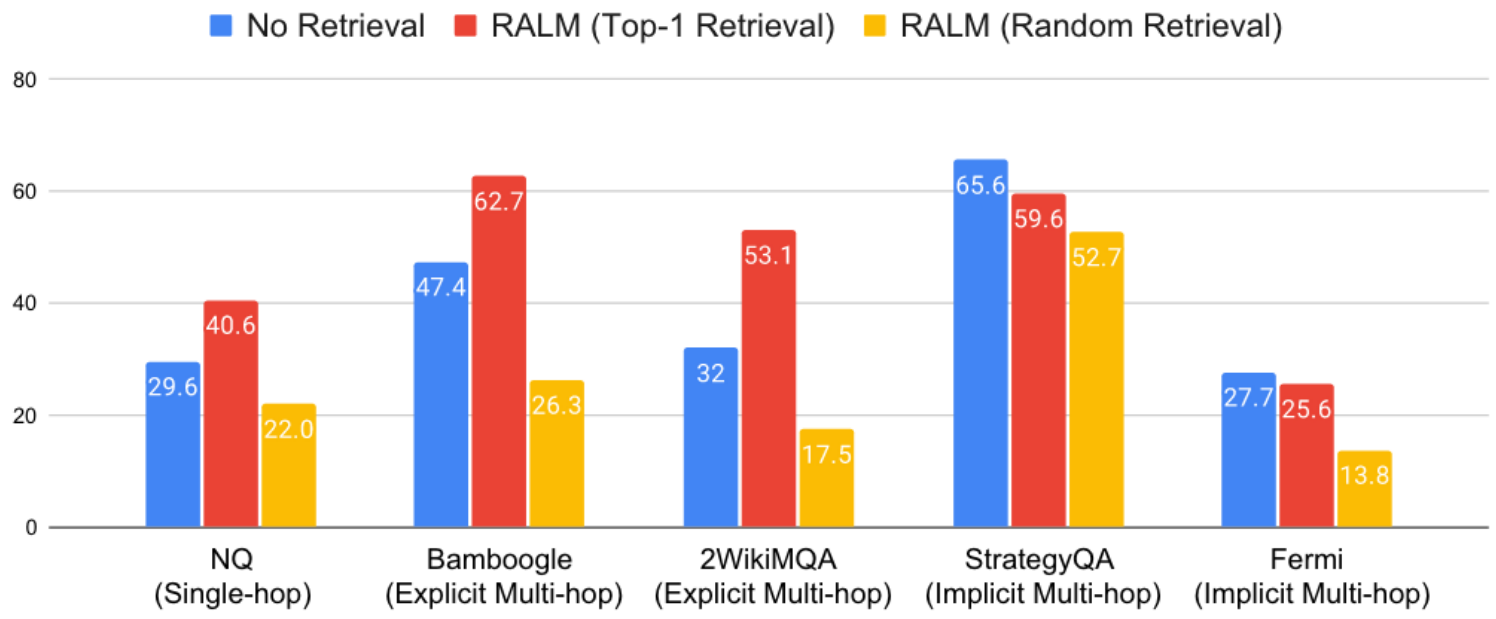


Search Augmented Instruction Learning; In-Context Pretraining: Language Modeling Beyond Document Boundaries

思考

AI Box

- 检索一定能增强大模型吗?
  - 无关信息可能会带来负面影响
  - 大模型已经掌握的知识不需要再检索

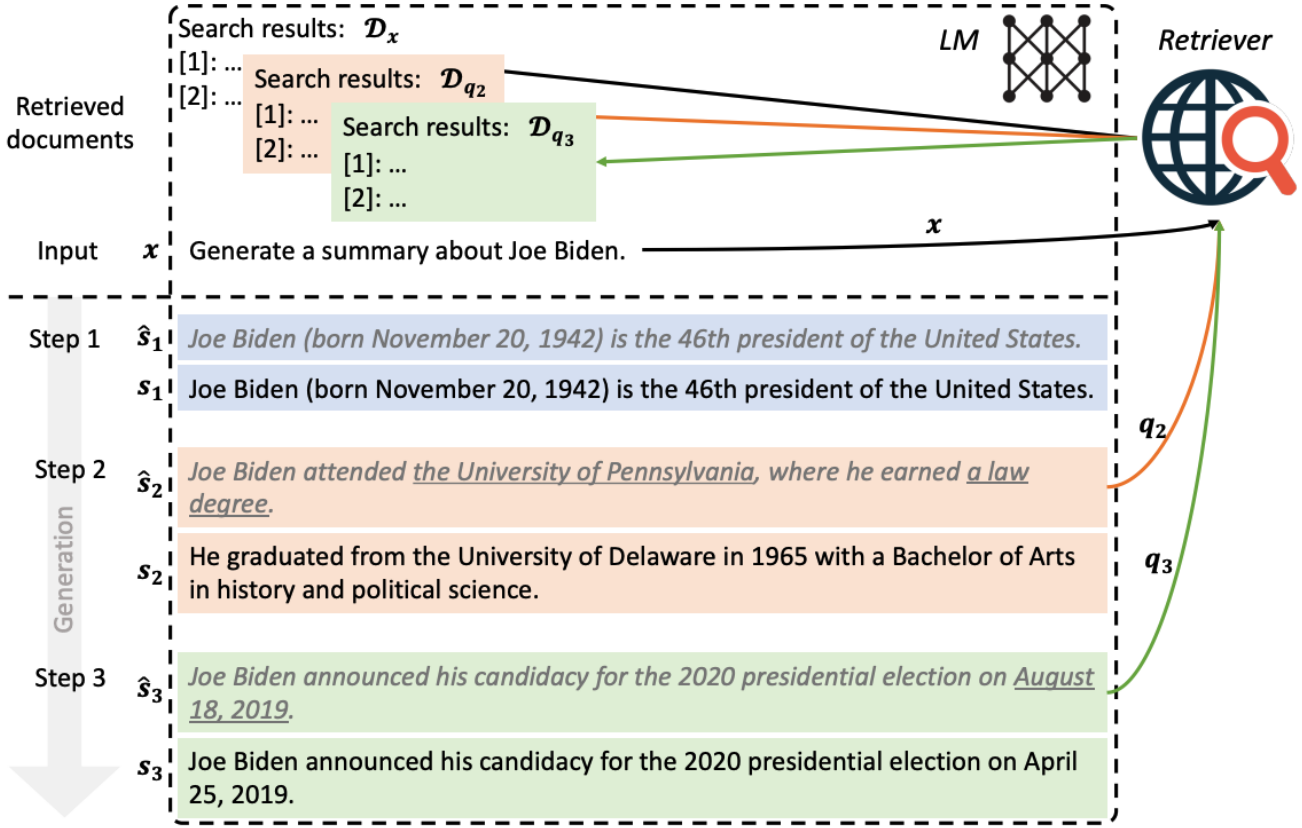


Making Retrieval-Augmented Language Models Robust to Irrelevant Context. ICLR 2024

思考

AI Box

- 如何判断是否需要检索？
  - 先生成后检索？
    - 可行，但成本高

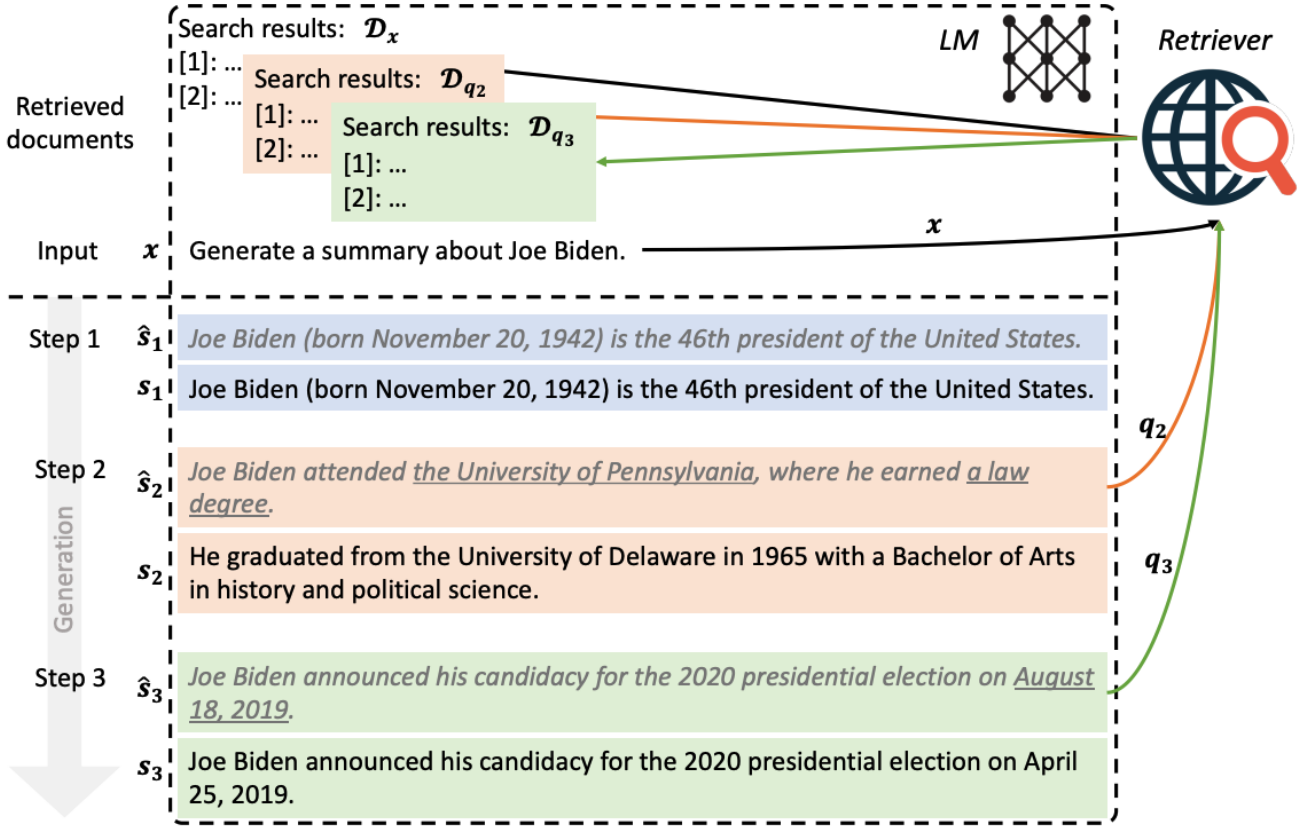


思考

AI Box

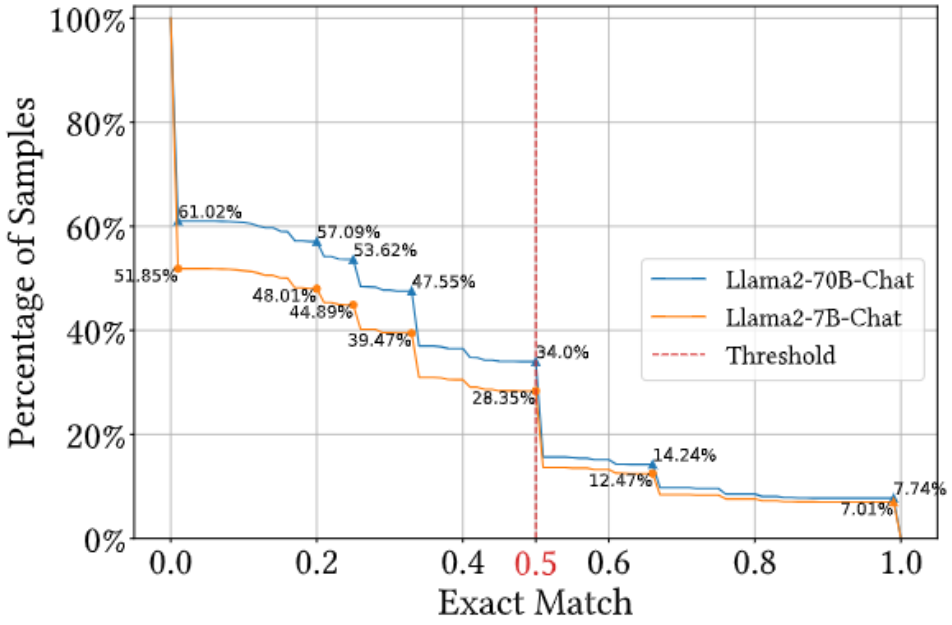
- 如何判断是否需要检索？
  - 先生成后检索？
    - 可行，但成本高

- 如何降低成本？
  - 用小模型进行生成并判断
  - 大小模型在知识程度上表现出相似性



思考

- 目前的大模型预训练使用了类似的训练语料
  - 不同模型在掌握程度高的知识是相当重合的
  - 大小模型知识能力的差距主要体现在长尾知识上



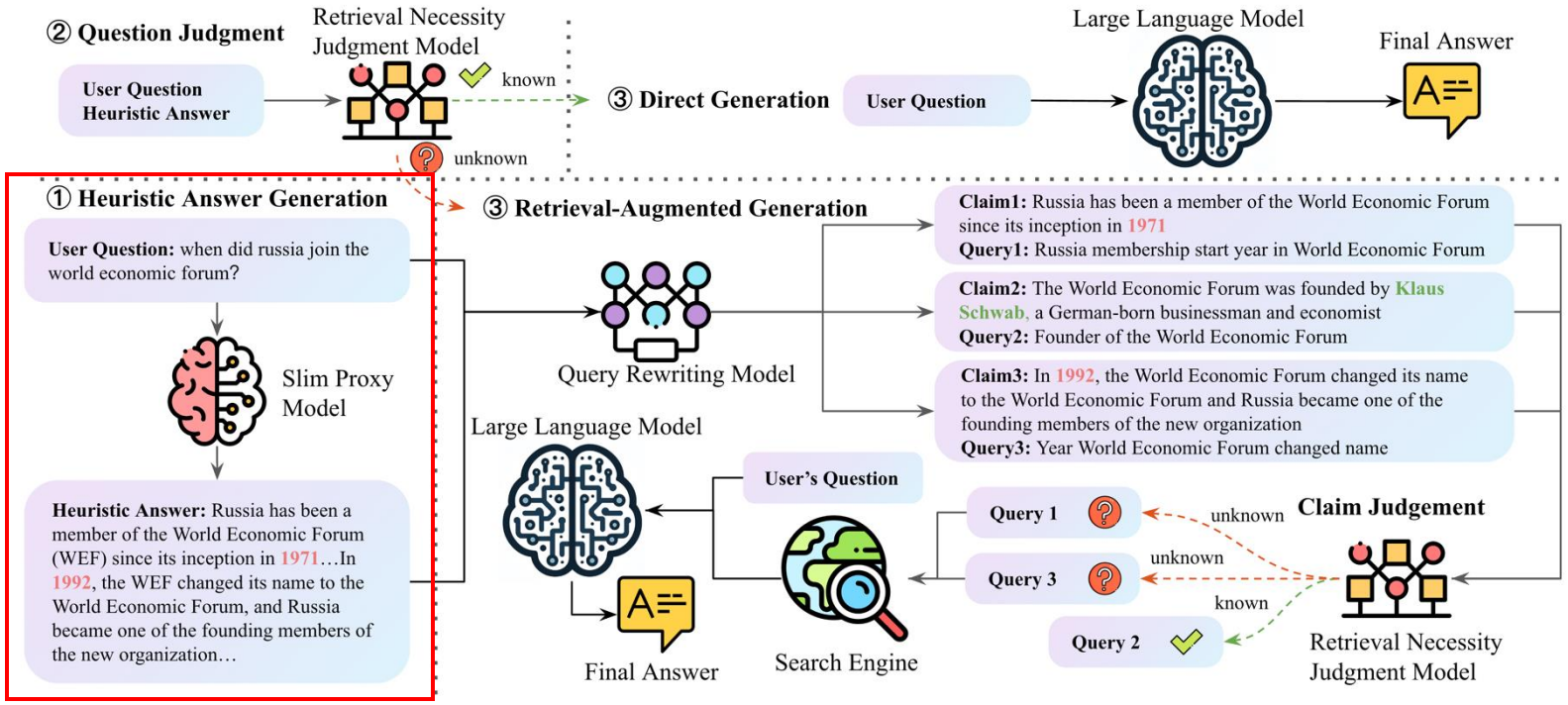
回答质量超过当前EM值的样本占总体样本的比例，越高说明当前模型能力越强

> 0.5 两条线迅速接近，两个模型 EM>0.5 以上的样本超过80%是一致的

基于代理模型的检索时机判定方法

AI Box

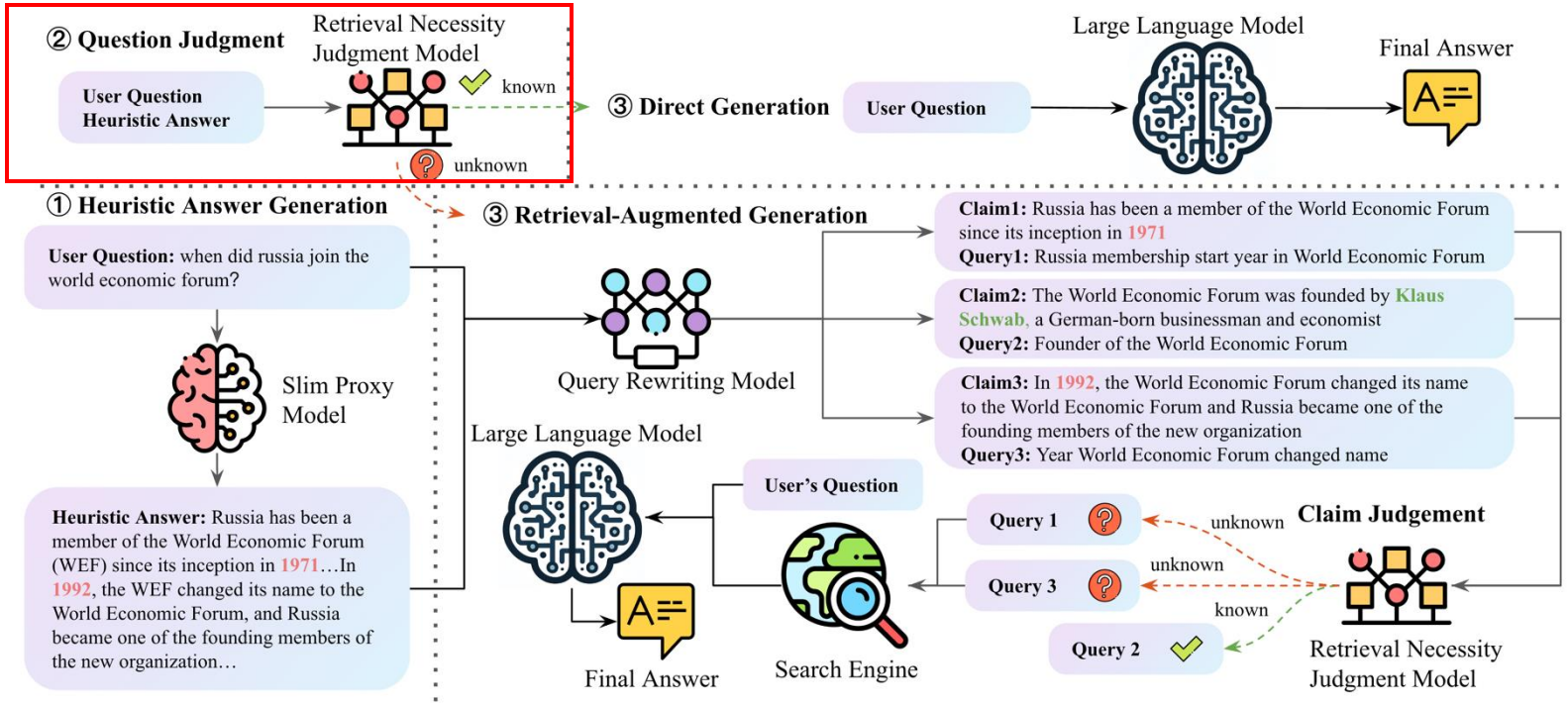
- 代理模型 (不做微调)
  - 生成初始回答 (Heuristic Answer)



基于代理模型的检索时机判定方法

AI Box

- 检索必要性判别模型 (微调后的 LLaMA-7B)
  - 决定当前问题是否需要检索、某个子查询是否检索

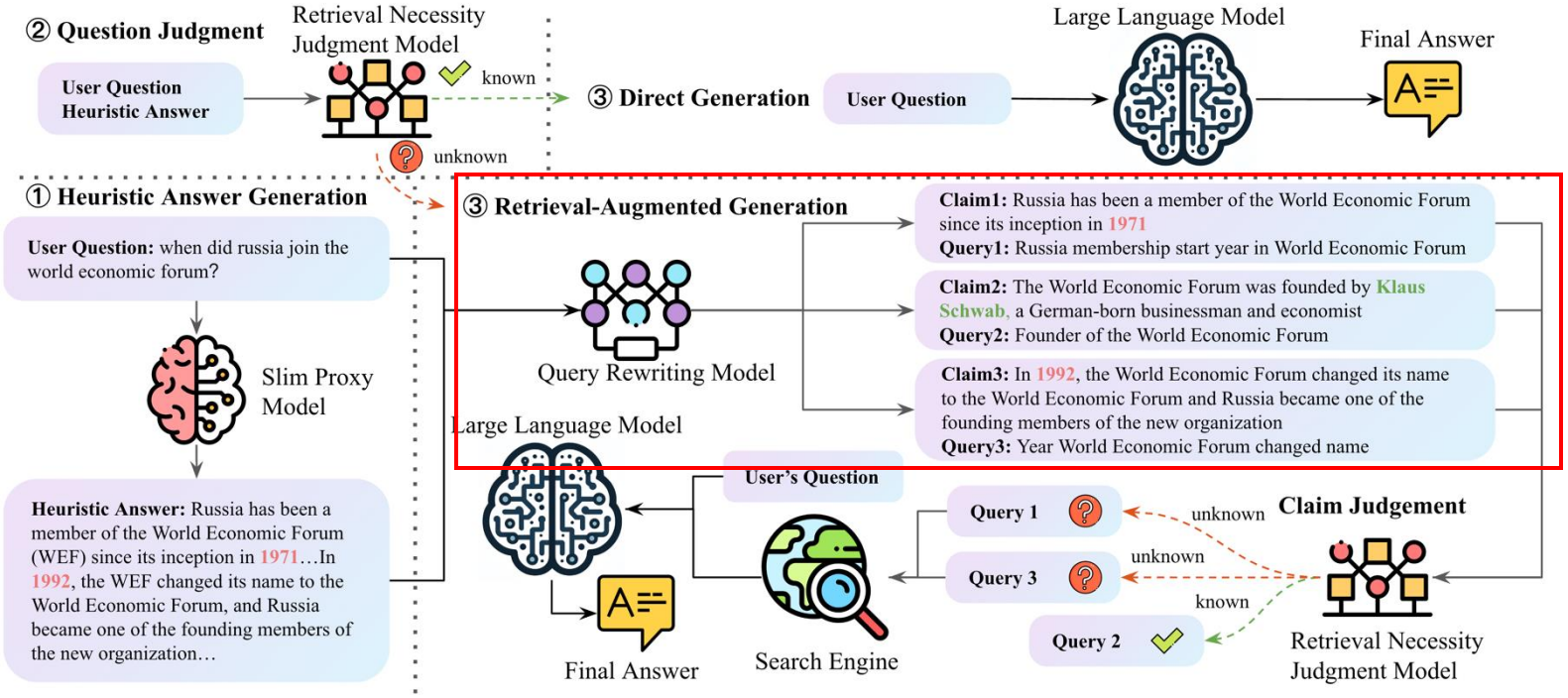


Small Models, Big Insights: Leveraging Slim Proxy Models to Decide When and What to Retrieve for LLMs. ACL 2024

基于代理模型的检索时机判定方法

AI Box

- 查询重写模型 (微调后的 LLaMA-7B)
  - 在当前问题需要检索的前提下，生成若干个子查询



基于代理模型的检索时机判定方法

AI Box

➢ 实验结果

- 相比于依赖大模型生成结果或概率的方法，SlimPLM 效果更好
- SlimPLM 只需要大模型推理一次，节省大量开销
- 与仅用用户输入进行判断的方法相比，SlimPLM 具备优势（代理模型解码结果具备指导性）

Method	#API	ASQA		NQ		Trivia-QA		MuSiQue	ELI5		
		EM	Hit@1	EM	Hit@1	EM	Hit@1	EM	ROUGE-1	ROUGE-2	ROUGE-L
Llama2-70B-Chat without Retrieval											
Vanilla Chat	1	29.68	62.50	40.49	55.00	27.44	90.75	11.50	28.66	4.88	14.27
CoT	1	26.21	54.50	35.36	48.75	23.50	79.00	11.50	28.12	4.73	14.06
Llama2-70B-Chat with Retrieval											
Direct RAG	1	27.63	58.00	42.40	56.00	28.07	92.25	10.50	28.61	4.76	**15.76**
FLARE	2.10	30.08	63.50	41.36	55.75	27.41	89.50	11.25	27.95	4.72	13.91
Self-Eval	2	29.45	60.75	42.15	55.75	27.58	91.50	10.25	28.70	4.83	15.39
Self-Ask	2.67	26.37	60.25	38.56	53.00	26.56	89.50	9.50	-	-	-
ITER-RETGEN	3	30.15	60.50	42.85	55.50	28.31	91.00	13.00	28.44	4.74	15.72
SKR-KNN	1	29.38	61.75	41.90	55.75	28.16	**92.25**	10.25	28.71	4.80	15.73
SlimPLM (Ours)	1	**30.73**	**65.00**	**47.43**	**62.25**	**28.35**	92.00	**13.00**	**29.97**	**5.61**	15.13
Qwen-72B-Chat without Retrieval											
Vanilla Chat	1	26.65	58.50	40.38	53.75	27.82	90.25	11.75	**30.61**	5.21	15.90
CoT	1	27.74	59.50	40.49	53.75	27.62	91.75	12.75	29.94	4.94	14.75
Qwen-72B-Chat with Retrieval											
Direct RAG	1	25.85	57.00	41.27	52.75	26.39	87.75	7.75	25.93	4.55	**16.74**
FLARE	2.29	27.68	59.00	40.89	54.50	27.10	88.50	**12.75**	30.31	5.20	15.77
Self-Eval	2	27.64	60.00	42.43	56.00	27.13	90.50	7.75	29.19	5.14	16.05
Self-Ask	2.76	22.82	52.25	36.16	49.25	25.29	87.50	9.75	-	-	-
ITER-RETGEN	3	**29.38**	61.50	43.51	57.50	27.16	89.75	12.25	26.15	4.41	16.52
SKR-KNN	1	28.08	61.50	43.08	56.00	26.38	88.50	11.25	27.29	4.75	16.31
SlimPLM (Ours)	1	27.97	**62.25**	**44.07**	**57.75**	**28.03**	**92.25**	9.75	29.56	**5.91**	16.36

Small Models, Big Insights: Leveraging Slim Proxy Models to Decide When and What to Retrieve for LLMs. ACL 2024

谢谢