检索增强的LLM

报告人：傅延赫

报告时间：2024.01.05

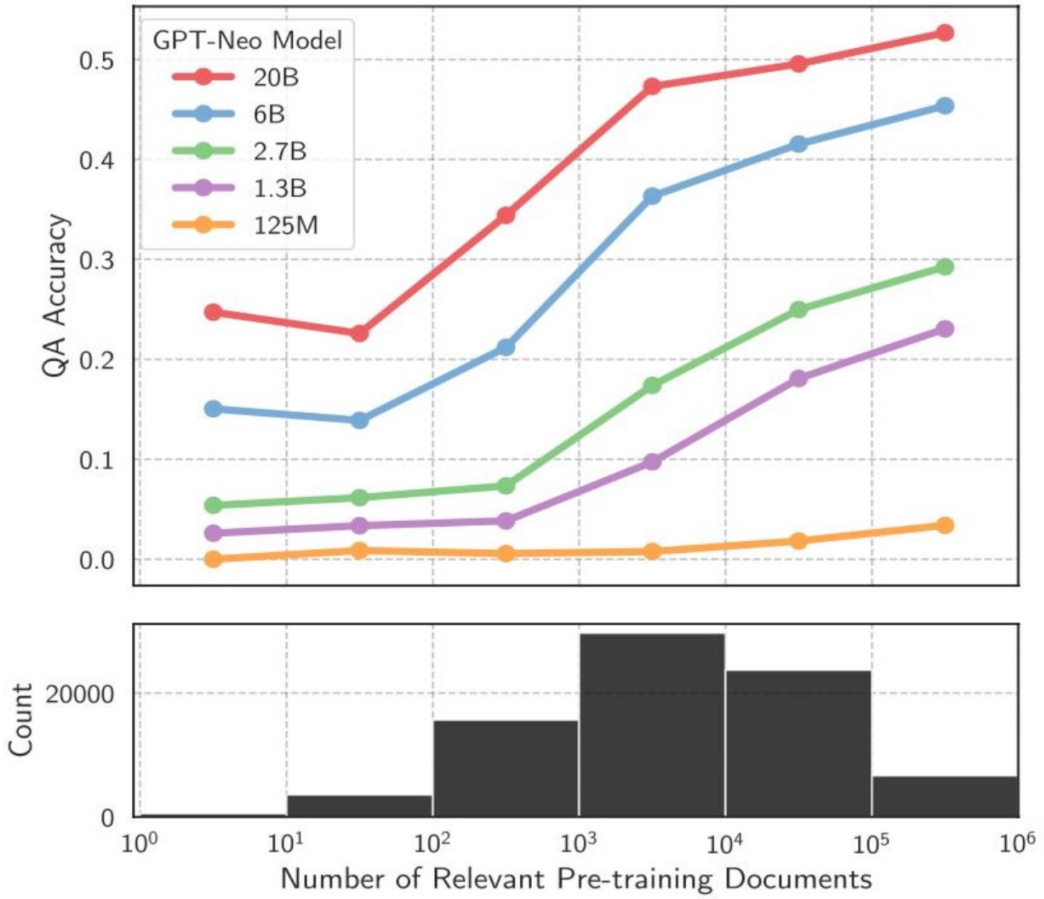
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

目录

目录

LLM的问题

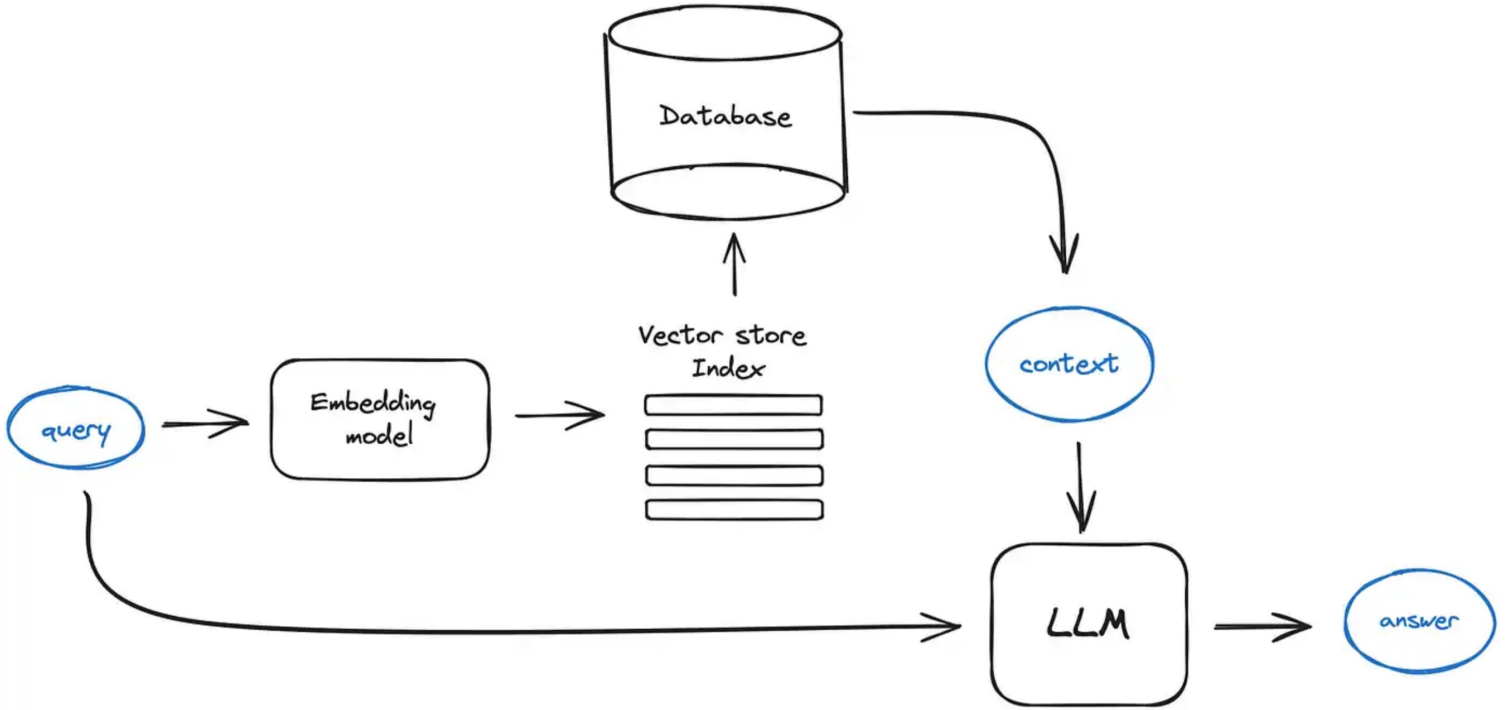
- 幻觉问题
- 长尾问题
- 私有数据
- 信息更新
- 可解释性



Large Language Models Struggle to Learn Long-Tail Knowledge, ICML 2023

检索增强生成（RAG）

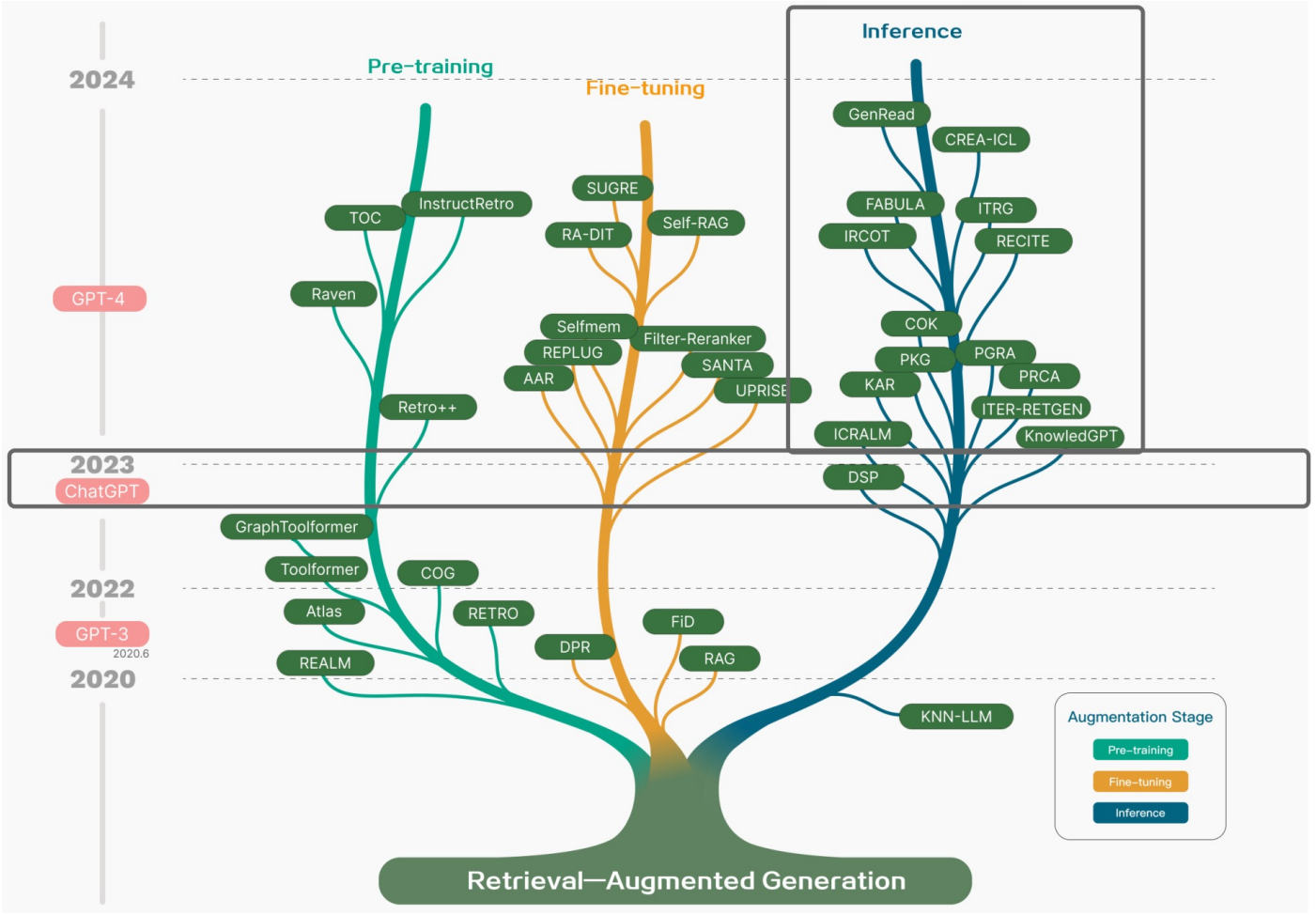
- 定义：先基于用户查询检索相关文档，然后将用户查询和文档一起输入给LLM完成回复
  - 半参数化方法，弥补参数化过程中带来的损失
  - 直观、有效、正交的问题解决方案



https://pub.towardsai.net/advanced-rag-techniques-an-illustrated-overview-04d193d8fec6

RAG发展

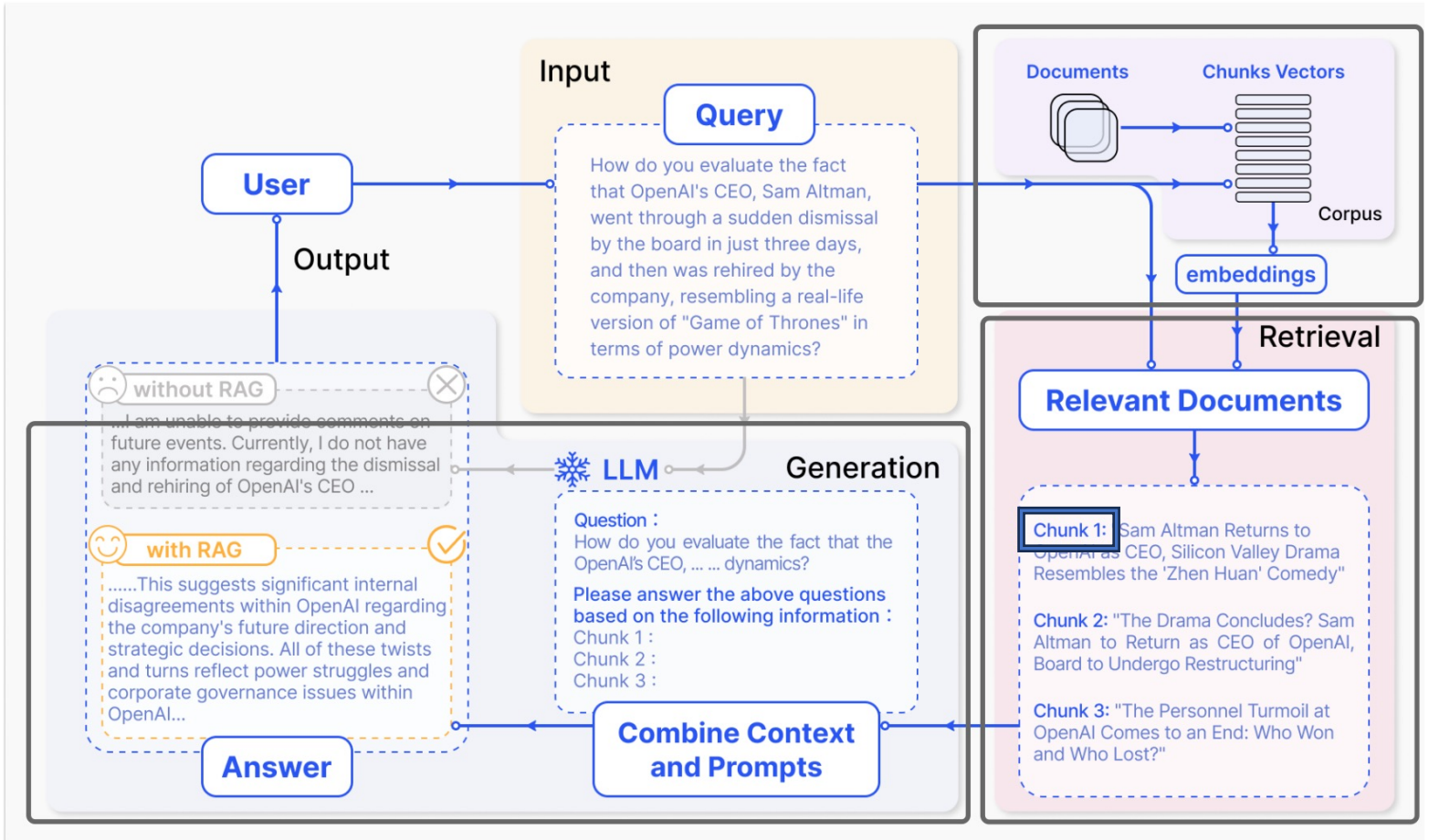
1. 最早提出者：Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks，NIPS 2020，Facebook
2. ChatGPT后，RAG的相关工作量大幅度提升
3. Inference阶段的工作量涨幅巨大，大量工作聚集于该阶段



Retrieval-Augmented Generation for Large Language Models: A Survey

RAG基本组成

- 由索引、检索、生成三个部分顺序组成。一个例子：

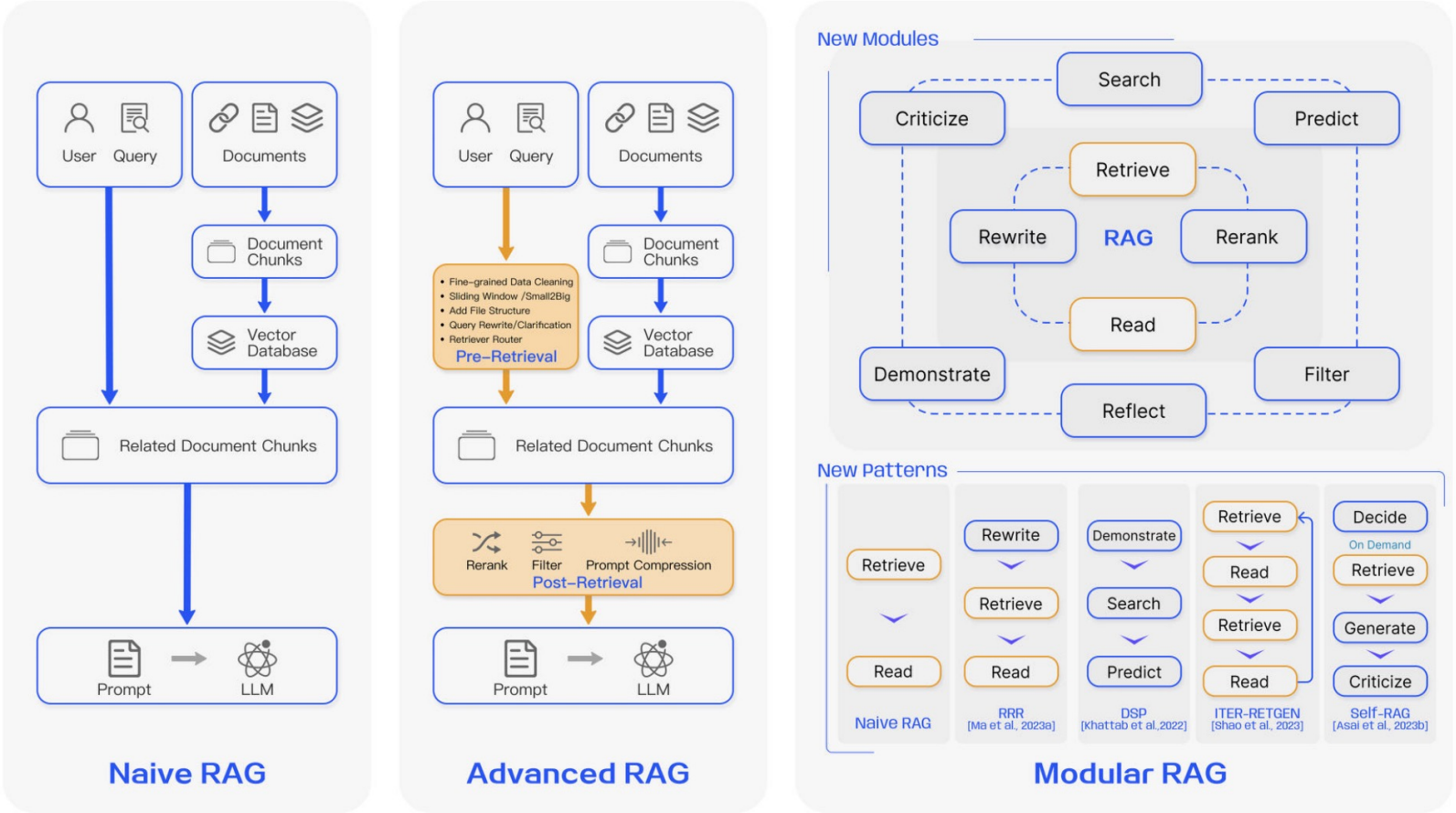


Retrieval-Augmented Generation for Large Language Models: A Survey

研究&改进方向

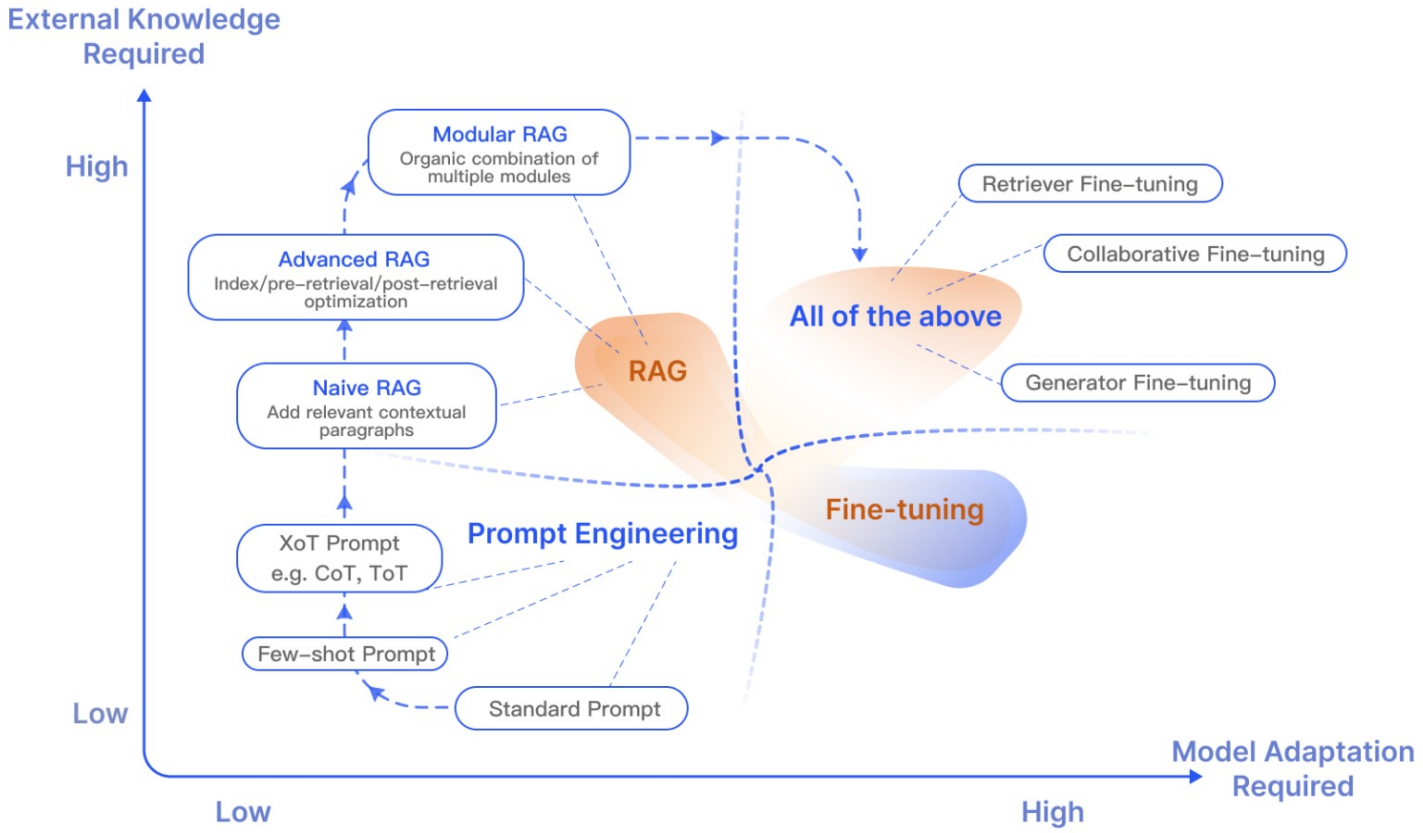
- Advanced RAG：聚焦预处理和后处理

- Modular RAG：
  - 已经打破了索引、检索、生成的基本组成架构
  - 新模块（工业界）
  - 新范式（学术界）



RAG与微调

- RAG给模型提供一本参考书，让LLM根据特定的问题去查找信息增加全新的知识
- 微调让模型通过深入持久的学习来吸收知识，突出模型参数中已有的知识，教会它理解广泛的领域或学习新的语言、格式或风格
- 二者相互补充



Retrieval-Augmented Generation for Large Language Models: A Survey

目录

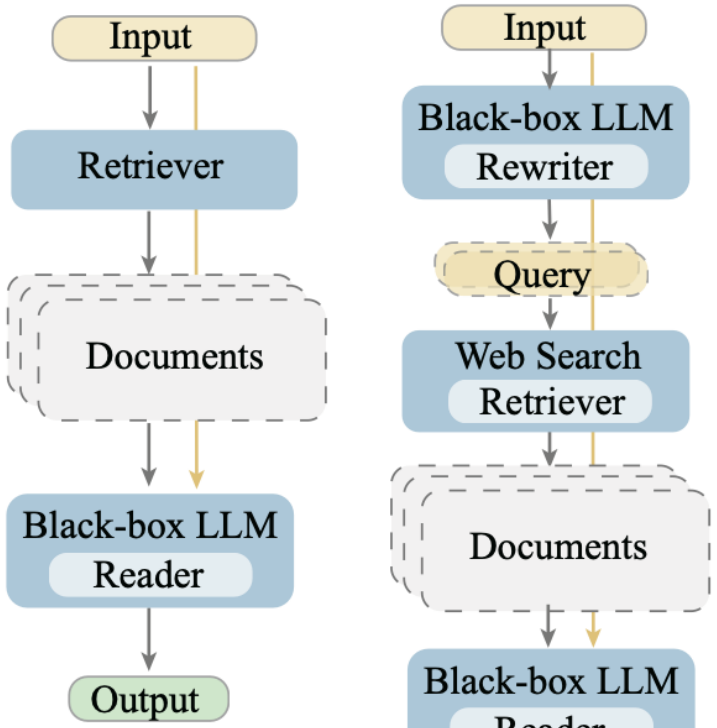
Rewrite-Retrieve-Read (RRR)

Query Rewriting for Retrieval-Augmented Large Language Models

Xinbei Ma[1,2,*] , **Yeyun Gong**[3, #, †], **Pengcheng He**[4, #], **Hai Zhao**[1,2,†], **Nan Duan**[3]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University
[3]Microsoft Research Asia [4]Microsoft Azure AI
sjtumaxb@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn,
{yegong, nanduan}@microsoft.com, Herbert.he@gmail.com

动机

- 最近的RAG方法将LLM作为Reader，并以适配LLM为导向进行设计：
  1. 训练一个Retriever来迎合LLM，使用LLM的反馈作为训练目标，检索模型可根据LLM的输入语境进行调整；
  2. 优化Retriever和Reader之间的交互行为，二者都是冻结的，通过设计prompt或复杂的管道来最大化发挥RAG的能力。
- 问题：忽略了用户输入和真正需要查询的知识之间的gap，限制了检索性能，给提示工程带来了负担。
- 本文提出：Rewrite-Retrieve-Read框架，在检索器前面增加了一个重写输入的步骤。转化用户输入更适合Retriever。



(a) Retrieve-then-read (b)Rewrite-retrieve-read

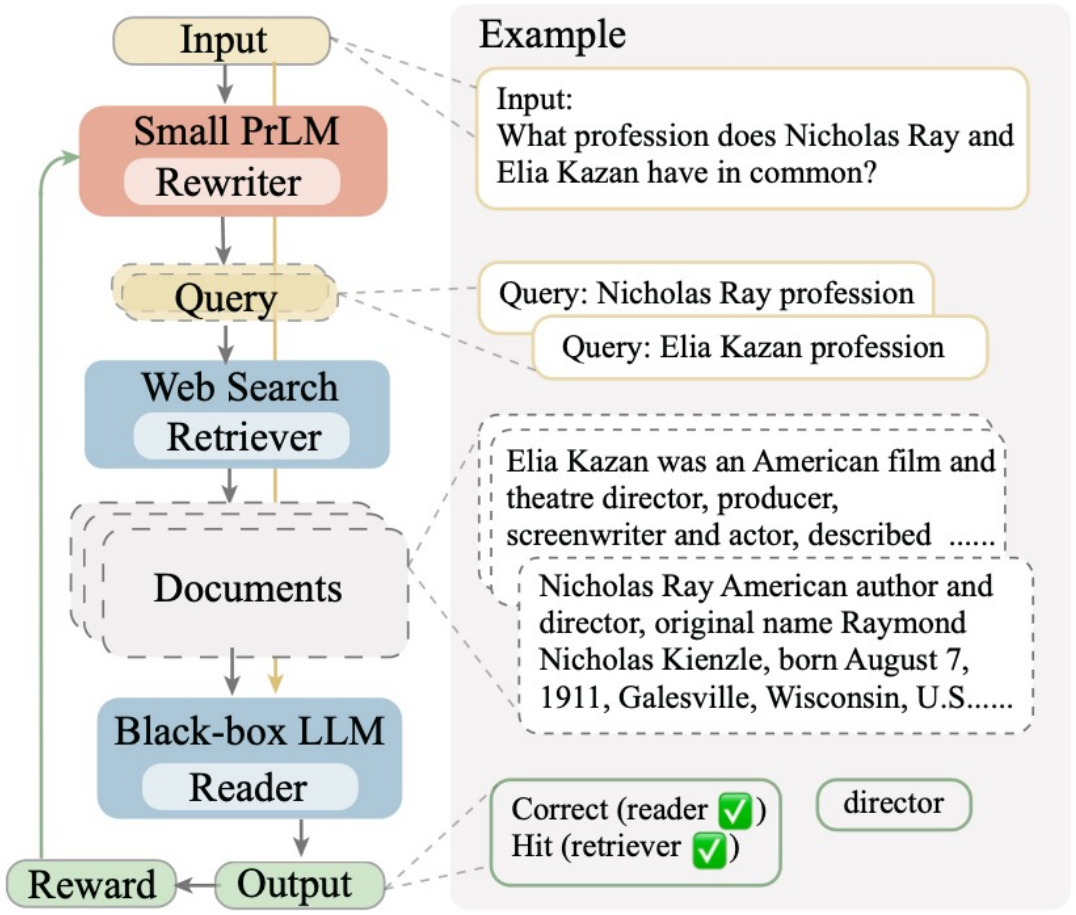
Query Rewriting for Retrieval-Augmented Large Language Models, EMNLP2023

方法

- 冻结Retriever和Reader，训练一个小模型作为Rewriter重写用户输入
- 训练过程：
  1. 预热（蒸馏）：让LLM重写数据集中的原始问题，然后筛选其中Reader可以做对的问题构成数据集，直接finetune；
  2. 强化学习：
     1. 预热后的模型是初始化policy model
     2. Action是生成下一个token $\hat{\tilde{x}}_t$
     3. 当前state是 $s_t = [x, \hat{\tilde{x}}_{<t}]$
     4. 生成重写结果后，根据reader和retriever结果得到reward

$$R_{lm} = EM + \lambda_f F_1 + \lambda_h Hit.$$

$$R(s_t, a_t) = R_{lm}(\hat{\tilde{x}}, y) - \beta \mathrm{KL}(\pi_\theta \| \pi_0).$$



(c) Trainable rewrite-retrieve-read

实验

- Rewriter：LLM或者FT的T5
- Retriever：Bing搜索引擎
- Reader：gpt-3.5-turbo或Vicuna-13B
- 数据集1：三个QA数据集
  - HotPotQA：需要多跳推理，实验结果表明了使用复杂问题作为查询不能弥补参数化知识的不足，反而会带来噪音。
  - AmbigNQ
  - PopQA
- 数据集 2：多选数据集MMLU

Model	EM	F_1
HotpotQA		
Direct	32.36	43.05
Retrieve-then-read	30.47	41.34
LLM rewriter	32.80	43.85
Trainable rewriter	34.38	45.97
AmbigNQ		
Direct	42.10	53.05
Retrieve-then-read	45.80	58.50
LLM rewriter	46.40	58.74
Trainable rewriter	47.80	60.71
PopQA		
Direct	41.94	44.61
Retrieve-then-read	43.20	47.53
LLM rewriter	46.00	49.74
Trainable rewriter	45.72	49.51

MMLU	EM			
	Human.	STEM	Other	Social
ChatGPT				
Direct	75.6	58.8	69.0	71.6
Retrieve-then-read	76.7	63.3	70.0	78.2
LLM rewriter	77.0	63.5	72.6	76.4
Vicuna-13B				
Direct	39.8	34.9	50.2	46.6
Retrieve-then-read	40.2	39.8	55.2	50.6
LLM rewriter	42.0	41.5	57.1	52.2
Trainable rewriter	43.2	40.9	59.3	51.2

Case Study

Example 1: multi-hop question	Hit	Correct
Q0: The youngest daughter of Lady Mary-Gaye Curzon stars with Douglas Smith and Lucien Laviscount in what 2017 film?	❌	❌
Q1: the youngest daughter of Lady Mary-Gaye Curzon; 2017 film stars Douglas Smith and Lucien Laviscount	✅	✅
Q2: Lady Mary-Gaye Curzon youngest daughter 2017 film with Douglas Smith and Lucien Laviscount	✅	✅
Example 2:		
Q0: What 2000 movie does the song "All Star" appear in?	❌	❌
Q1: movie "All Star" 2000	❌	❌
Q2: 2000 movie "All Star" song	✅	✅

Example 3: multiple choice		
Q0: A car-manufacturing factory is considering a new site for its next plant. Which of the following would community planners be most concerned with before allowing the plant to be built? Options: A. The amount of materials stored in the plant B. The hours of operations of the new plant C. The effect the plant will have on the environment D. The work environment for the employees at the plant	❌	❌
Q1: What would community planners be most concerned with before allowing a car-manufacturing factory to be built?	✅	✅

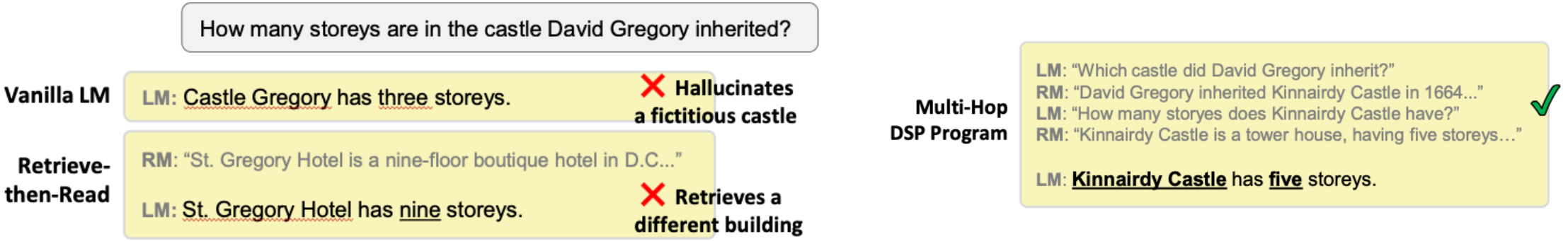
Figure 3: Examples for intuitive illustration. Q0 denotes original input, Q1 is from the LLM rewriter, and Q2 is from the trained T5 rewriter. **Hit** means retriever recall the answer, while **Correct** is for the reader output.

Iterative Retrieval-Generation (ITER-RETGEN)

Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy

Zhihong Shao[1], Yeyun Gong[2], yelong shen[3], Minlie Huang[1]*, Nan Duan[2], Weizhu Chen[3]
[1] The CoAI Group, DCST, Institute for Artificial Intelligence,
[1] State Key Lab of Intelligent Technology and Systems,
[1] Beijing National Research Center for Information Science and Technology,
[1] Tsinghua University, Beijing 100084, China
[2] Microsoft Research Asia [3] Microsoft Azure AI
szh19@mails.tsinghua.edu.cn aihuang@tsinghua.edu.cn

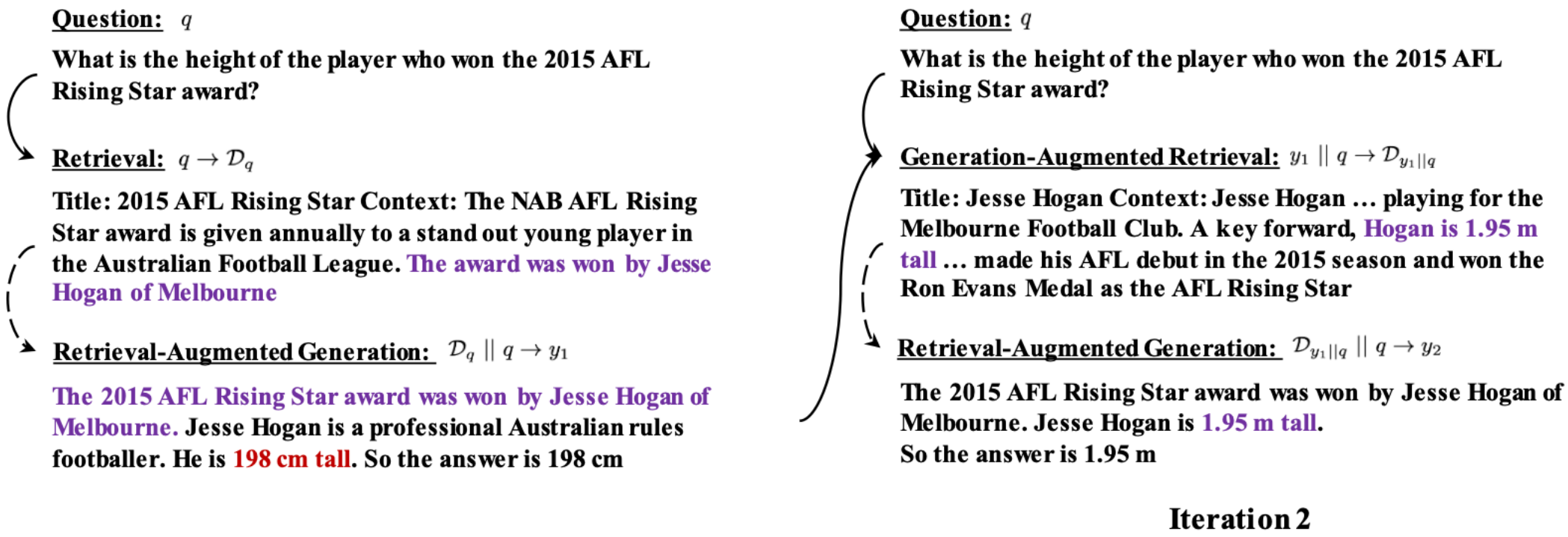
动机



- 以前的RAG采用一次性检索，如果用户输入中的需求很明确，那么一次性检索足够，但这不适用于需求复杂的任务，例如多跳推理和长文本问答，因为检索结果可能不可靠。
- ITER-RETGEN迭代·检索增强生成和·生成增强检索。不同于其他迭代模型，ITER-RETGEN的检索过程不会中断生成，而是利用上一次的完整生成结果来检索更多相关信息；这使ITER-RETGEN还可以在生成过程中处理所有检索到的知识。

方法

- 检索增强生成：用Dense Retriever和蒸馏后的Re-ranker配合思维链格式的prompt完成生成。
- 生成增强检索：用上一轮的生成结果和输入检索，来弥补Retriever的表示瓶颈导致的语义空白。



实验

Datasets	Example
HotPotQA	What is the name of this American musician, singer, actor, comedian, and songwriter, who worked with Modern Records and born in December 5, 1932?
2WikiMultiHopQA	Which film came out first, Blind Shaft or The Mask Of Fu Manchu?
MuSiQue	In which year did the publisher of In Cold Blood form?
Bamboogle	When did the first prime minister of the Russian Empire come into office?
Feverous	Is it true that Based on the same platform as the Chevrolet Sail, the Baojun 310 was launched on 2017 Beijing Auto Show where the price ranges from 36.800 yuan to 60.800 yuan?
StrategyQA	Is it common to see frost during some college commencements?

- baselines：
  - ReAct：重复推理、行动和观察。推理就是CoT生成；行动可以是生成查询以检索，也可以是最终确定答案；观察是连接检索段落；
  - Self-Ask：在回答每个问题之前会问“Are follow up questions needed here:”如果模型自己生成yes，就自己提一个问题，然后RAG这个问题。从而一步步回答最初问题；
  - DSP：由多跳检索阶段和答案生成阶段组成。在检索阶段的每一跳，都会提示模型生成搜索查询，并总结检索知识以供后续使用。在预测阶段，DSP 根据总结的知识和检索到的文档，使用 CoT 生成答案。

实验

Method	HotPotQA			2WikiMultiHopQA			MuSiQue			Bamboogle			Feverous		StrategyQA	
	EM	F1	Acc†	EM	F1	Acc†	EM	F1	Acc†	EM	F1	Acc†	Acc	Acc†	Acc	Acc†
Without Retrieval																
Direct	21.9	36.8	44.8	21.3	29.2	33.9	7.0	18.7	15.8	11.2	24.4	28.0	60.1	60.1	66.5	66.7
CoT	30.0	44.1	50.0	30.0	39.6	44.0	19.4	30.9	28.6	**43.2**	**51.1**	60.0	59.8	59.8	71.0	71.0
With Retrieval																
Direct	31.6	44.7	53.3	27.3	35.4	43.6	13.9	28.2	26.5	17.6	31.8	43.2	69.8	69.8	65.6	65.6
ReAct	24.9	44.7	61.1	28.0	38.5	45.9	23.4	37.0	37.9	21.8	31.0	40.3	66.4	66.4	66.9	66.9
Self-Ask	36.8	55.2	64.8	**37.3**	**48.8**	55.9	**27.6**	41.5	**42.9**	31.5	41.2	54.8	70.7	70.7	70.2	70.2
DSP	43.8	55.0	60.8	-	-	-	-	-	-	-	-	-	-	-	-	-
ITER-RETGEN 1	39.2	53.9	65.5	33.7	45.2	55.4	24.2	38.6	38.1	36.8	47.7	57.6	67.0	67.0	72.0	72.0
ITER-RETGEN 2	44.1	58.6	71.2	34.9	47.0	58.1	26.4	41.1	41.0	38.4	48.7	59.2	68.8	68.8	73.0	73.0
ITER-RETGEN 3	45.2	59.9	71.4	34.8	47.8	58.3	25.7	41.4	40.8	37.6	47.0	59.2	69.0	69.0	72.3	72.3
ITER-RETGEN 4	45.8	**61.1**	**73.4**	36.0	47.4	58.5	26.7	41.8	40.8	38.4	49.6	60.0	**71.5**	**71.5**	73.8	73.8
ITER-RETGEN 5	45.2	60.3	72.8	35.5	47.5	58.8	25.7	40.7	39.6	39.2	49.7	**60.8**	70.3	70.3	73.2	73.2
ITER-RETGEN 6	**45.9**	61.0	73.3	35.5	48.1	**59.4**	25.9	40.5	39.8	40.0	50.0	59.2	70.9	70.9	72.4	72.4
ITER-RETGEN 7	45.1	60.4	72.9	35.5	47.4	58.4	26.1	**42.0**	41.0	40.0	50.7	**60.8**	70.5	70.5	**74.1**	**74.1**

Table 2: Evaluation results on multi-hop question answering, fact verification, and commonsense reasoning datasets. Acc† is the accuracy of model outputs evaluated with `text-davinci-003`. For ITER-RETGEN, we evaluated LLM outputs in different iterations (up to 7 iterations). Underlined metric values are higher than those of Self-Ask.

Case Study

HotPotQA

Question: The arena where the Lewiston Maineiacs played their home games can seat how many people?
Answer: 3,677 seated

Iteration 1

Retrieved Knowledge

(1) Title: Lewiston Maineiacs Context: The Lewiston Maineiacs were a junior ice hockey team ... The team played its home games at the Androscoggin Bank Colisée ...
(2) Title: Bangor Auditorium Context: ... a 5,948 seat multi-purpose arena located in Bangor, Maine ...

Generation

The Lewiston Maineiacs were a junior ice hockey team of the Quebec Major Junior Hockey League based in Lewiston, Maine. The team played its home games at the Androscoggin Bank Colisée. The Androscoggin Bank Colisée can seat up to 5,948 people. So the answer is 5,948

Iteration 2

Retrieved Knowledge

(1) Title: Lewiston Maineiacs Context: The Lewiston Maineiacs were a junior ice hockey team ... The team played its home games at the Androscoggin Bank Colisée ...
(2) Title: Androscoggin Bank Colisée Context: The Androscoggin Bank Colisée (formerly Central Maine Civic Center and Lewiston Colisee) is a 4,000 capacity (3,677 seated) multi-purpose arena ...

Generation

The Lewiston Maineiacs were a junior ice hockey team of the Quebec Major Junior Hockey League based in Lewiston, Maine. They played their home games at the Androscoggin Bank Colisée. The Androscoggin Bank Colisée has a capacity of 4,000 (3,677 seated). So the answer is 3,677

Self-RAG

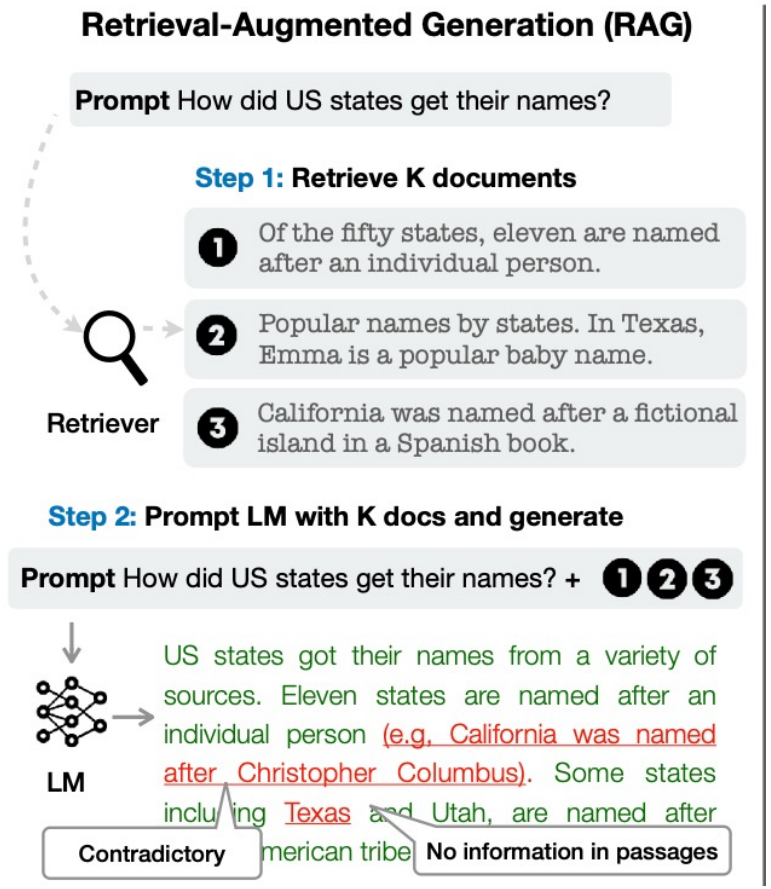
SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION

Akari Asai[†], **Zeqiu Wu**[†], **Yizhong Wang**[†§], **Avirup Sil**[‡], **Hannaneh Hajishirzi**[†§]
[†]University of Washington [§]Allen Institute for AI [‡]IBM Research AI
{akari,zeqiuwu,yizhongw,hannaneh}@cs.washington.edu, avi@us.ibm.com

动机

- 传统RAG不加区分地检索段落，导致检索结果可能没帮助或有噪音。此外，由于模型没有经过明确的训练如何使用检索结果中的知识，因此无法保证输出结果与检索到的文档保持一致。

- 本文引入自我反思的RAG机制，使LLM学会在生成输出的过程中，自适应地检索段落（即：模型可以判断是否要进行RAG），还引入了反思令牌（reflection tokens），使LM在推理阶段可控。



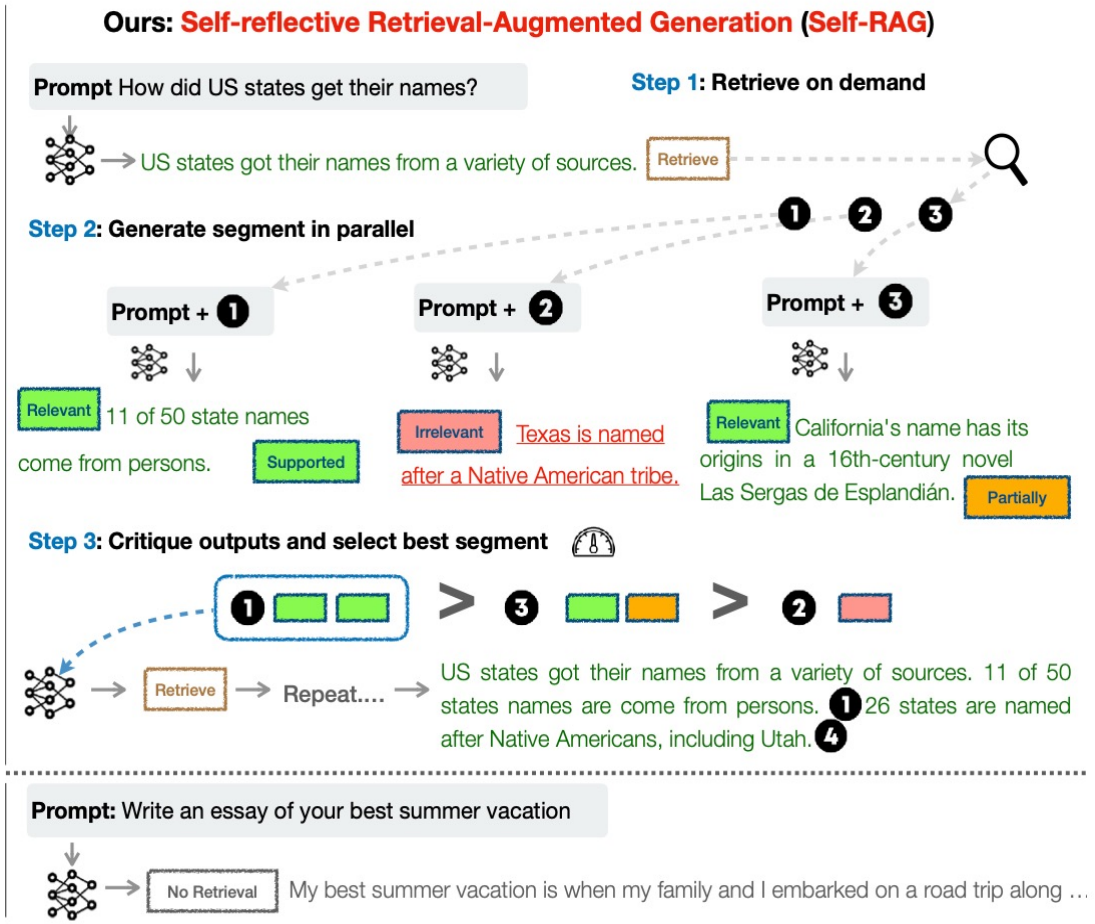
方法（Inference）

Type	Input	Output	Definitions
Retrieve	x / x, y	{yes, no, continue}	Decides when to retrieve with $\mathcal{R}$
IsRel	x, d	{**relevant**, irrelevant}	d provides useful information to solve x.
IsSup	x, d, y	{**fully supported**, partially supported, no support}	All of the verification-worthy statement in y is supported by d.
IsUse	x, y	{**5**, 4, 3, 2, 1}	y is a useful response to x.

Table 1: Four types of reflection tokens used in SELF-RAG. Each type uses several tokens to represent its output values. The bottom three rows are three types of Critique tokens, and **the bold text** indicates the most desirable critique tokens. x, y, d indicate input, output, and a relevant passage, respectively.

Algorithm 1 SELF-RAG Inference

Require: Generator LM $\mathcal{M}$, Retriever $\mathcal{R}$, Large-scale passage collections $\{d_1, \ldots, d_N\}$
Input: input prompt x and preceding generation $y_{<t}$, **Output:** next output segment y_t
$\mathcal{M}$ predicts Retrieve given $(x, y_{<t})$
if Retrieve == Yes **then**
  Retrieve relevant text passages $\mathbf{D}$ using $\mathcal{R}$ given (x, y_{t-1}) ▷ Retrieve
  $\mathcal{M}$ predicts IsRel given x, d and y_t given $x, d, y_{<t}$ for each $d \in \mathbf{D}$ ▷ Generate
  $\mathcal{M}$ predicts IsSup and IsUse given x, y_t, d for each $d \in \mathbf{D}$ ▷ Critique
  Rank y_t based on IsRel , IsSup , IsUse ▷ Detailed in Section 3.3
else if Retrieve == No **then**
  $\mathcal{M}_{gen}$ predicts y_t given x ▷ Generate
  $\mathcal{M}_{gen}$ predicts IsUse given x, y_t ▷ Critique



方法（Training）

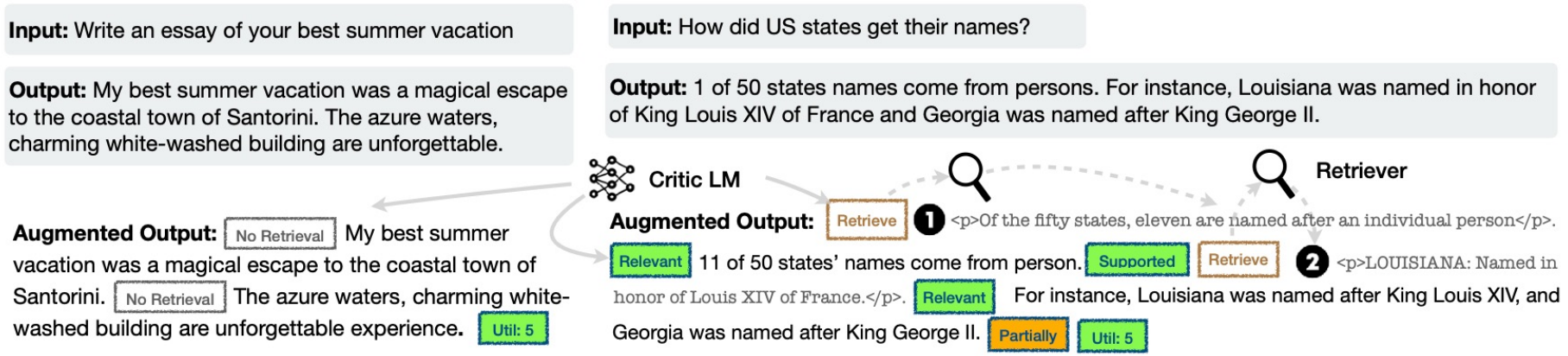


Figure 2: SELF-RAG training examples. The left example does not require retrieval while the right one requires retrieval; thus, passages are inserted. More examples are in Appendix Table 4.

1. 用ICL + ChatGPT 为文本标上四类特殊标签，用这些数据训练一个Critic LM
2. 用Critic LM在纯文本中的指定位置插入四类标签，使每条文本数据包含特殊标签token
3. 用2中的数据直接训练LLM

- 注：需要扩充词表

实验

LM	Short-form PopQA (acc)	Short-form TQA (acc)	Closed-set Pub (acc)	Closed-set ARC (acc)	Long-form generations (with citations) Bio (FS)	ASQA (em)	ASQA (rg)	ASQA (mau)	ASQA (pre)	ASQA (rec)
LMs with proprietary data										
Llama2-c$_{13B}$	20.0	59.3	49.4	38.4	55.9	22.4	29.6	28.6	–	–
Ret-Llama2-c$_{13B}$	51.8	59.8	52.1	37.9	79.9	32.8	34.8	43.8	19.8	36.1
ChatGPT	29.3	74.3	70.1	75.3	71.8	35.3	36.2	68.8	–	–
Ret-ChatGPT	50.8	65.7	54.7	75.3	–	40.7	39.9	79.7	65.1	76.6
Perplexity.ai	–	–	–	–	71.2	–	–	–	–	–
Baselines without retrieval										
Llama2$_{7B}$	14.7	30.5	34.2	21.8	44.5	7.9	15.3	19.0	–	–
Alpaca$_{7B}$	23.6	54.5	49.8	45.0	45.8	18.8	29.4	61.7	–	–
Llama2$_{13B}$	14.7	38.5	29.4	29.4	53.4	7.2	12.4	16.0	–	–
Alpaca$_{13B}$	24.4	61.3	55.5	54.9	50.2	22.9	32.0	70.6	–	–
CoVE$_{65B}$ *	–	–	–	–	71.2	–	–	–	–	–
Baselines with retrieval										
Toolformer*$_{6B}$	–	48.8	–	–	–	–	–	–	–	–
Llama2$_{7B}$	38.2	42.5	30.0	48.0	78.0	15.2	22.1	32.0	2.9	4.0
Alpaca$_{7B}$	46.7	64.1	40.2	48.0	76.6	30.9	33.3	57.9	5.5	7.2
Llama2-FT$_{7B}$	48.7	57.3	64.3	65.8	78.2	31.0	35.8	51.2	5.0	7.5
SAIL*$_{7B}$	–	–	69.2	48.4	–	–	–	–	–	–
Llama2$_{13B}$	45.7	47.0	30.2	26.0	77.5	16.3	20.5	24.7	2.3	3.6
Alpaca$_{13B}$	46.1	66.9	51.1	57.6	77.7	**34.8**	36.7	56.6	2.0	3.8
Our SELF-RAG $_{7B}$	54.9	66.4	72.4	67.3	**81.2**	30.0	35.7	**74.3**	66.9	67.8
Our SELF-RAG $_{13B}$	**55.8**	**69.3**	**74.5**	**73.1**	80.2	31.7	**37.0**	71.6	**70.3**	**71.3**

QA、多选、传记生成

训练数据不开源的模型（RAG和非RAG）

非RAG的LLMs

RAG的LLMs

Case Study

Input When did sunday became the seventh day of the week (asqa)

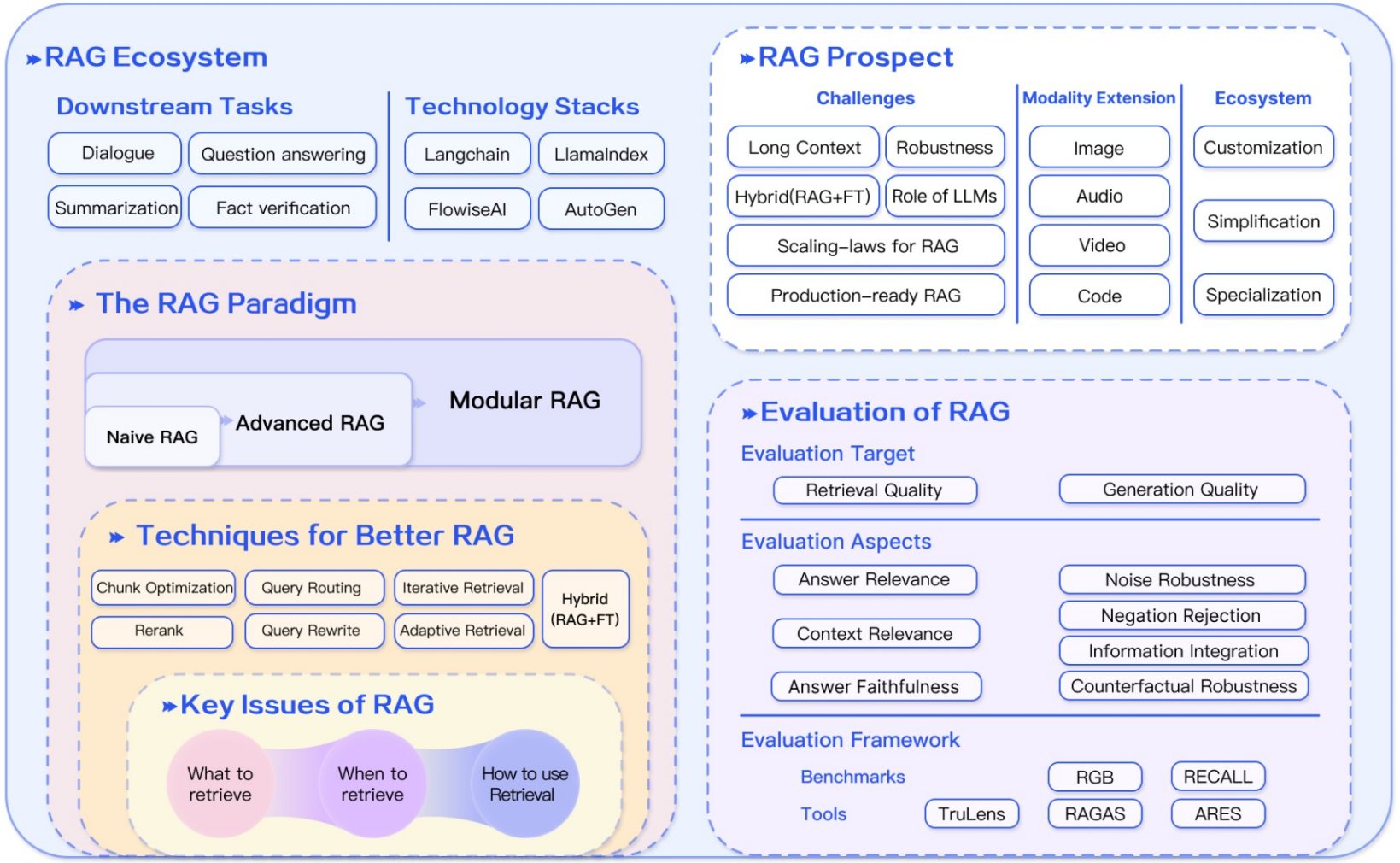
Output [Retrieve=Yes] <p>Reference [1]<p>[IsRel =Relevant] In 321 AD, the Roman Emperor Constantine the Great made Sunday a day of rest from labor, stating: "On the venerable Day of the Sun let the magistrates and people residing in cities rest, and let all workshops be closed".[IsSup =Fully Supported] [Retrieve=Yes] <p>Reference [2]<p>[IsRel =Relevant] The official adoption of Sunday as a day of rest by Constantine the Great in AD 321 made Sunday the seventh day of the week.**[2]**[IsSup =Fully Supported][IsUse =5]

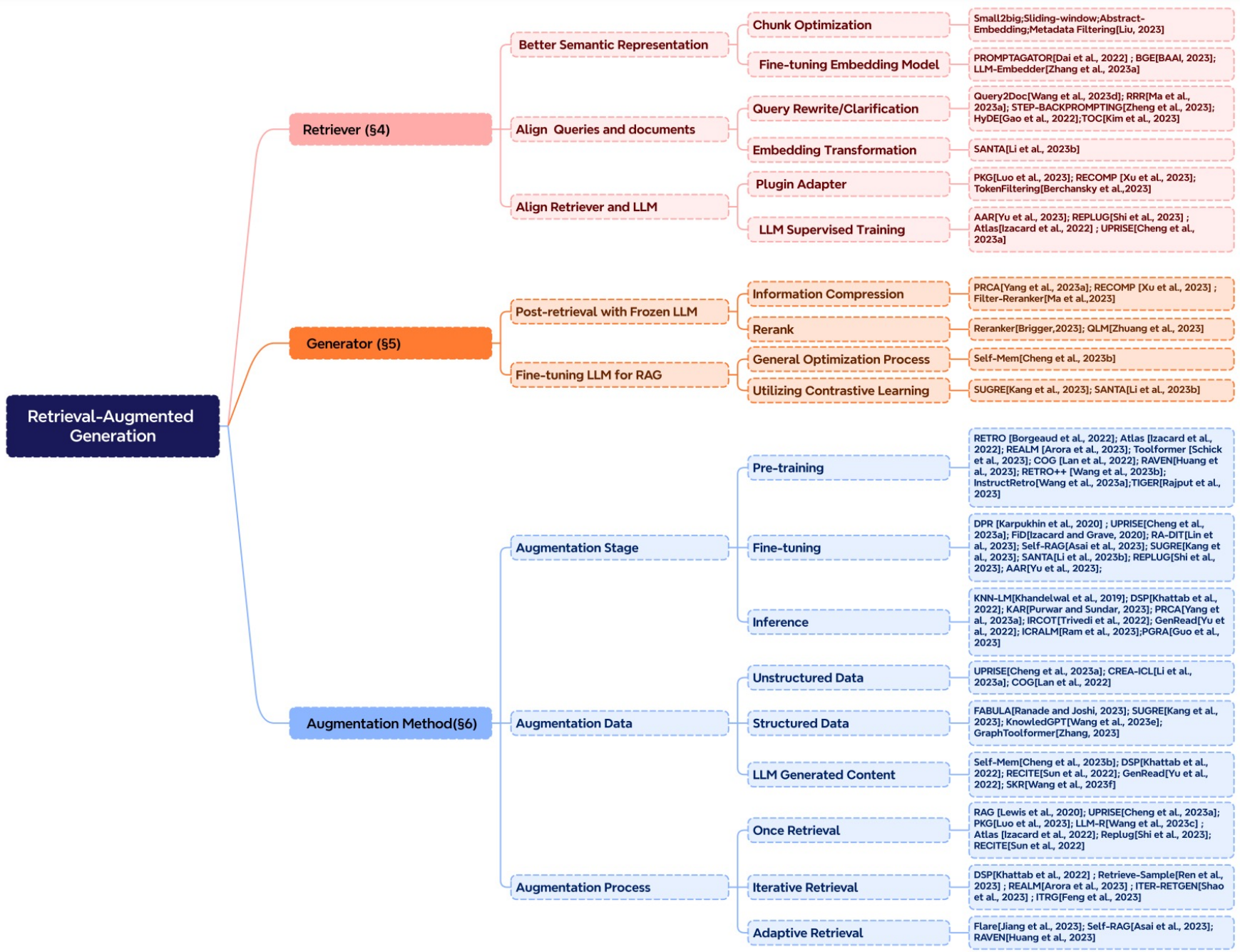
Reference [1] Emperor Constantine then enacted the first Sunday Laws, for "the venerable Day of the Sun" in 321 A.D. On March 7, 321, the Roman emperor Constantine issued a decree making Sunday a day of rest from labor, stating: We find a good description of this in sources such as the book Rest Days which states: Early Christian observance of both the spiritual seventh-day sabbath and a Lordś Day assembly is evidenced in Ignatiusś letter to the Magnesians ca. 110.

Reference [2] Despite the official adoption of Sunday as a day of rest by Constantine, the seven-day week and the nundial cycle continued to be used side-by-side until at least the Calendar of 354 and probably later. ... The fact that the canon had to be issued at all is an indication that adoption of Constantine's decree of 321 was still not universal ...

目录

生态


RAG Ecosystem
Downstream Tasks
Dialogue
Question answering
Summarization
Fact verification
Technology Stacks
Langchain
LlamaIndex
FlowiseAI
AutoGen
The RAG Paradigm
Naive RAG
Advanced RAG
Modular RAG
Techniques for Better RAG
Chunk Optimization
Query Routing
Iterative Retrieval
Hybrid (RAG+FT)
Rerank
Query Rewrite
Adaptive Retrieval
Key Issues of RAG
What to retrieve
When to retrieve
How to use Retrieval
RAG Prospect
Challenges
Long Context
Robustness
Hybrid(RAG+FT)
Role of LLMs
Scaling-laws for RAG
Production-ready RAG
Modality Extension
Image
Audio
Video
Code
Ecosystem
Customization
Simplification
Specialization
Evaluation of RAG
Evaluation Target
Retrieval Quality
Generation Quality
Evaluation Aspects
Answer Relevance
Context Relevance
Answer Faithfulness
Noise Robustness
Negation Rejection
Information Integration
Counterfactual Robustness
Evaluation Framework
Benchmarks
RGB
RECALL
Tools
TruLens
RAGAS
ARES


Retrieval-Augmented Generation
Retriever (§4)
Better Semantic Representation
Chunk Optimization
Small2big;Sliding-window;Abstract-Embedding;Metadata Filtering[Liu, 2023]
Fine-tuning Embedding Model
PROMPTAGATOR[Dai et al., 2022] ; BGE[BAAI, 2023]; LLM-Embedder[Zhang et al., 2023a]
Align Queries and documents
Query Rewrite/Clarification
Query2Doc[Wang et al., 2023d]; RRR[Ma et al., 2023a]; STEP-BACKPROMPTING[Zheng et al., 2023]; HyDE[Gao et al., 2022];TOC[Kim et al., 2023]
Embedding Transformation
SANTA[Li et al., 2023b]
Align Retriever and LLM
Plugin Adapter
PKG[Luo et al., 2023]; RECOMP [Xu et al., 2023]; TokenFiltering[Berchansky et al.,2023]
LLM Supervised Training
AAR[Yu et al., 2023]; REPLUG[Shi et al., 2023] ; Atlas[Izacard et al., 2022] ; UPRISE[Cheng et al., 2023a]
Generator (§5)
Post-retrieval with Frozen LLM
Information Compression
PRCA[Yang et al., 2023a]; RECOMP [Xu et al., 2023] ; Filter-Reranker[Ma et al.,2023]
Rerank
Reranker[Brigger,2023]; QLM[Zhuang et al., 2023]
Fine-tuning LLM for RAG
General Optimization Process
Self-Mem[Cheng et al., 2023b]
Utilizing Contrastive Learning
SUGRE[Kang et al., 2023]; SANTA[Li et al., 2023b]
Augmentation Method(§6)
Augmentation Stage
Pre-training
RETRO [Borgeaud et al., 2022]; Atlas [Izacard et al., 2022]; REALM [Arora et al., 2023]; Toolformer [Schick et al., 2023]; COG [Lan et al., 2022]; RAVEN[Huang et al., 2023]; RETRO++ [Wang et al., 2023b]; InstructRetro[Wang et al., 2023a];TIGER[Rajput et al., 2023]
Fine-tuning
DPR [Karpukhin et al., 2020] ; UPRISE[Cheng et al., 2023a]; FiD[Izacard and Grave, 2020]; RA-DIT[Lin et al., 2023]; Self-RAG[Asai et al., 2023]; SUGRE[Kang et al., 2023]; SANTA[Li et al., 2023b]; REPLUG[Shi et al., 2023]; AAR[Yu et al., 2023];
Inference
KNN-LM[Khandelwal et al., 2019]; DSP[Khattab et al., 2022]; KAR[Purwar and Sundar, 2023]; PRCA[Yang et al., 2023a]; IRCOT[Trivedi et al., 2022]; GenRead[Yu et al., 2022]; ICRALM[Ram et al., 2023];PGRA[Guo et al., 2023]
Augmentation Data
Unstructured Data
UPRISE[Cheng et al., 2023a]; CREA-ICL[Li et al., 2023a]; COG[Lan et al., 2022]
Structured Data
FABULA[Ranade and Joshi, 2023]; SUGRE[Kang et al., 2023]; KnowledGPT[Wang et al., 2023e]; GraphToolformer[Zhang, 2023]
LLM Generated Content
Self-Mem[Cheng et al., 2023b]; DSP[Khattab et al., 2022]; RECITE[Sun et al., 2022]; GenRead[Yu et al., 2022]; SKR[Wang et al., 2023f]
Augmentation Process
Once Retrieval
RAG [Lewis et al., 2020]; UPRISE[Cheng et al., 2023a]; PKG[Luo et al., 2023]; LLM-R[Wang et al., 2023c] ; Atlas [Izacard et al., 2022]; Replug[Shi et al., 2023]; RECITE[Sun et al., 2022]
Iterative Retrieval
DSP[Khattab et al., 2022] ; Retrieve-Sample[Ren et al., 2023] ; REALM[Arora et al., 2023] ; ITER-RETGEN[Shao et al., 2023] ; ITRG[Feng et al., 2023]
Adaptive Retrieval
Flare[Jiang et al., 2023]; Self-RAG[Asai et al., 2023]; RAVEN[Huang et al., 2023]

Thanks!

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS