Parametric Retrieval Augmented Generation

李莹璐 2025.2.25

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

目录

- 研究现状
- 作者介绍
- 研究方法
- 实验效果

Parametric Retrieval Augmented Generation

Weihang Su
swh22@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing 100084, China

Yichen Tang*
DCST, Tsinghua University
Beijing 100084, China

Qingyao Ai†
aiqy@tsinghua.edu.cn
DCST, Tsinghua University
Beijing 100084, China

Junxi Yan
DCST, Tsinghua University
Beijing 100084, China

Changyue Wang
DCST, Tsinghua University
Beijing 100084, China

Hongning Wang
DCST, Tsinghua University
Beijing 100084, China

Ziyi Ye
DCST, Tsinghua University
Beijing 100084, China

Yujia Zhou
DCST, Tsinghua University
Beijing 100084, China

Yiqun Liu
DCST, Tsinghua University
Beijing 100084, China

Weihang Su

Other names ▸

Tsinghua University
Verified email at mails.tsinghua.edu.cn - Homepage

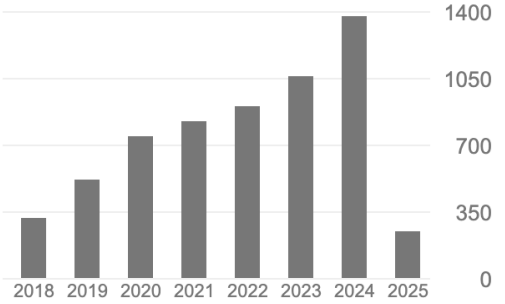
Information Retrieval Natural Language Processing AI for Legal

Qingyao Ai

Associate Professor, Dept. of CS&T, Tsinghua University
Verified email at tsinghua.edu.cn - Homepage

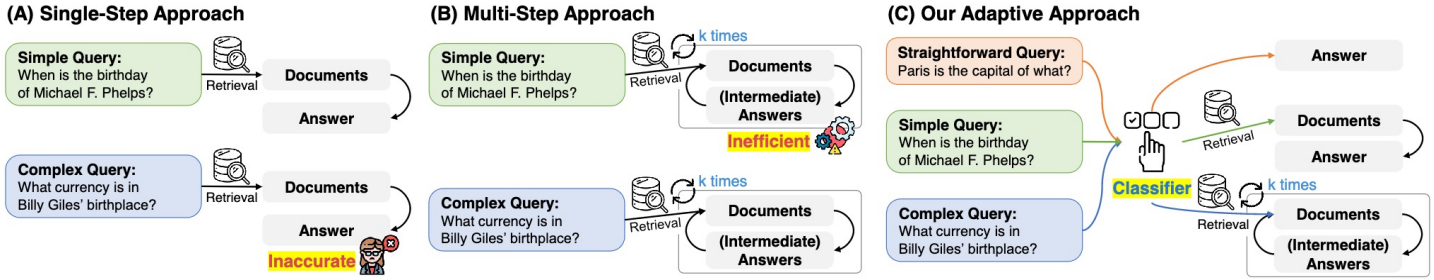
Information Retrieval Machine Learning

	All	Since 2020
Citations	6220	5186
h-index	31	30
i10-index	73	69

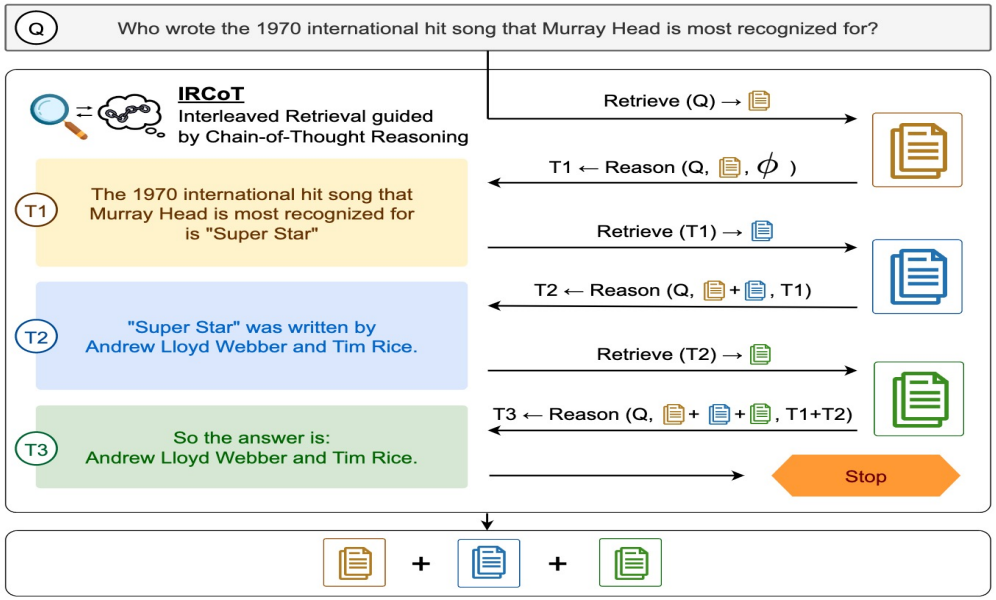


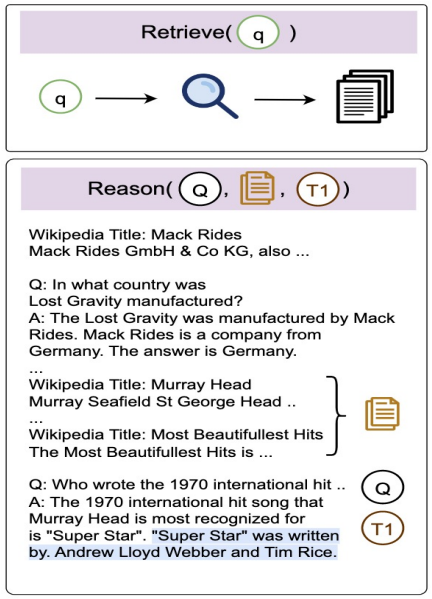
研究现状

- RAG框架扩展：包括考虑检索时机、知识库的构建、文档选择等方法。



AdapterRAG：
根据问题复杂程度自适应选择合适策略

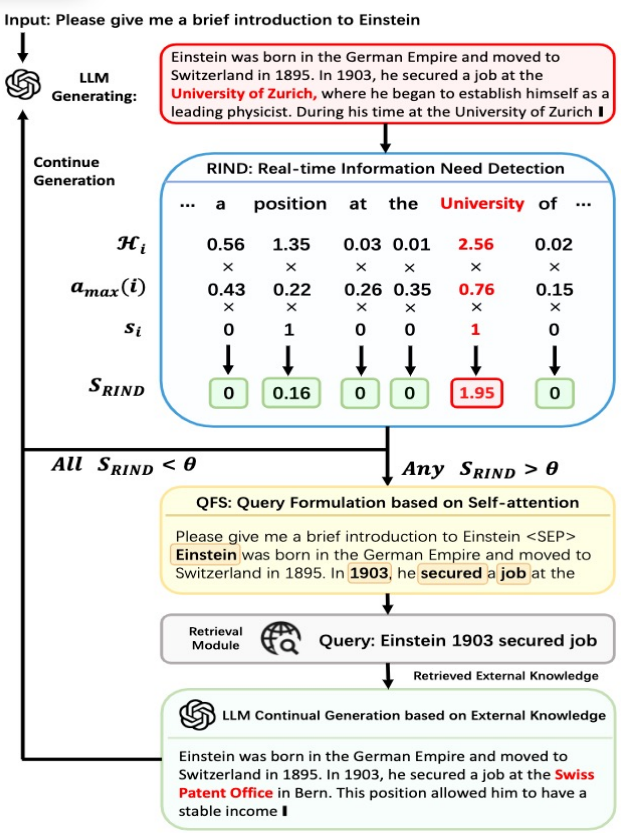




IRCoT：
交错的生成、检索CoT引导检索结果

研究现状

- 动态的RAG：主要针对长文本生成过程中信息需求变化的情况。



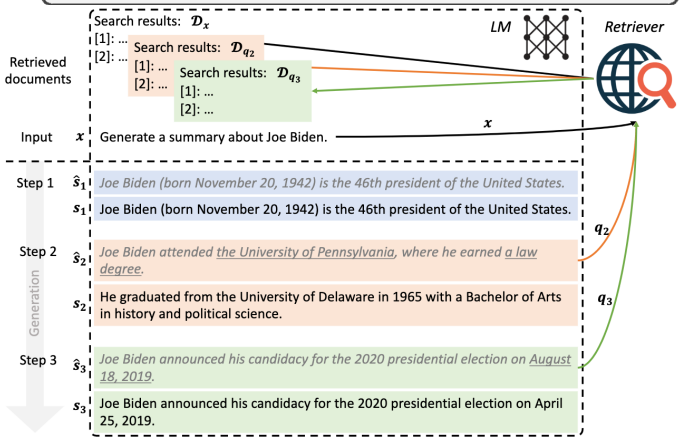
DRAGIN：
对LLM的实时信息需求进行建模

Prompt 3.1: retrieval instructions

Skill 1. An instruction to guide LMs to generate search queries.
Several search-related exemplars.

Skill 2. An instruction to guide LMs to perform a specific downstream task (e.g., multihop QA).
Several task-related exemplars.

An instruction to guide LMs to combine skills 1 and 2 for the test case.
The input of the test case.



FLARE：
根据模型的token预测概率改变检索重点内容

研究现状

- 现有的RAG大多是将检索到的相关知识文档附加到LLM的输入中来引导生成过程。
- 存在问题：
  - 增加上下文长度和相关文档数量会增大计算开销，并可能降低复杂推理任务的性能；
  - LLM主要在参数中存储知识，RAG主要在输入层面操作，限制了LLM利用外部知识的能力。
- 新范式：
  - 通过文档参数化直接将外部知识整合到LLM到前馈网络（FFN）中。

研究方法

- 离线文档参数化：将知识库中的每个文档都转换成plug-in参数
  - 文档增强
  - 参数化文档编码

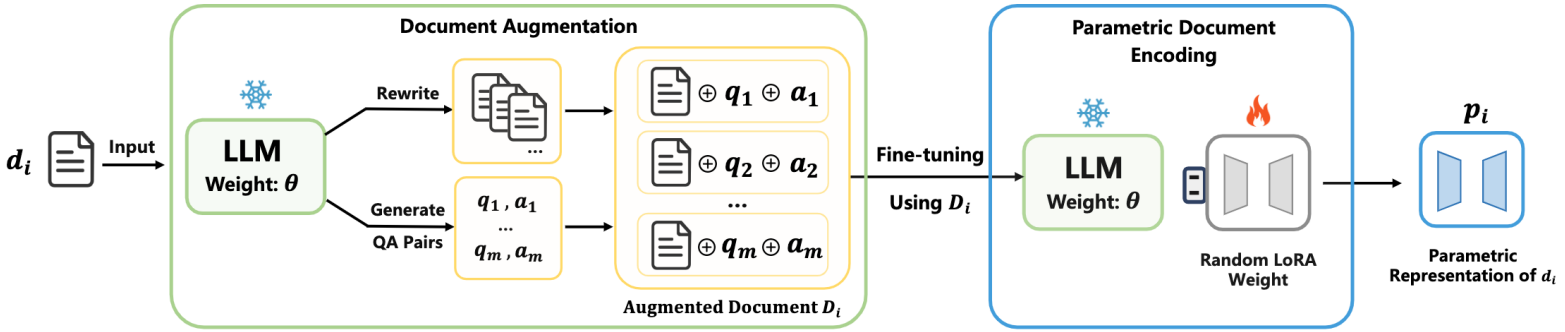


Figure 2: An illustration of how we parameterize each document d_i in the corpus during the *Offline Document Parameterization* stage.

研究方法

- 离线文档参数化：将知识库中的每个文档都转换成plug-in参数
  - 文档增强：通过文档改写和基于文档的问答对生成，将每个文档转换成包含多种语言变体的综合资源。

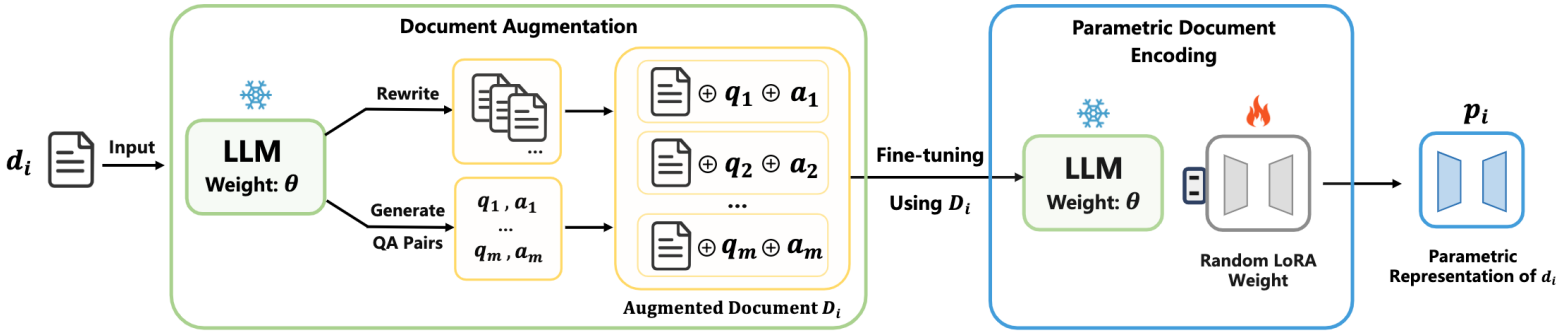


Figure 2: An illustration of how we parameterize each document d_i in the corpus during the *Offline Document Parameterization* stage.

研究方法

- 离线文档参数化：将知识库中的每个文档都转换成plug-in参数
  - 参数化文档编码：针对增强数据集的每个文档，提前计算出对应的lora。

$$W' = W + \Delta W = W + AB^{\top},$$

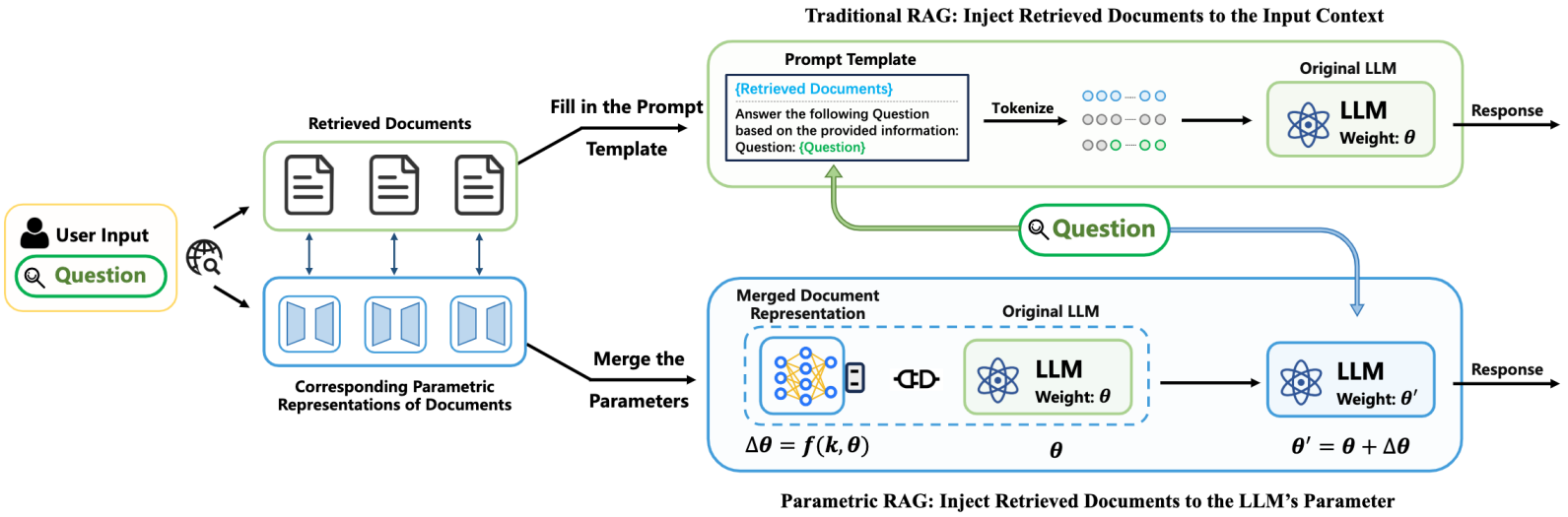
$$x = [\, d_i^{\,k} \oplus q_i^{\,j} \oplus a_i^{\,j} \,],$$

$$\min_{\Delta\theta} \sum_{(d_i^{\,k}, q_i^{\,j}, a_i^{\,j}) \in D_i} \sum_{t=1}^{T} -\log P_{\theta+\Delta\theta}(x_t \,|\, x_{<t}),$$

$$K_P = \{p_i \mid p_i = f_\phi(d_i), \quad i = 1, 2, \ldots, N\},$$

研究方法

- 在线推理：
  - 检索：使用检索器根据输入提示检索与查询最相关的文档；
  - 更新：将检索文档对应的参数与LLM原有参数合并，更新LLM的参数；
  - 生成：更新后的LLM根据原始输入提示生成回答。



实验效果

- 在一些开源LLM上优于现有的其他RAG方法

Table 1: The overall experiment results of Parametric RAG and other baselines across four tasks. All metrics reported are F1 scores. Bold numbers indicate the best performance of all baselines, and the second-best results are underlined. "*" and † denote significantly worse performance than the bolded method and our proposed P-RAG with $p < 0.05$ level, respectively.

		2WikiMultihopQA					HotpotQA			PopQA	CWQ
		Compare	Bridge	Inf.	Compose	Total	Bridge	Compare	Total		
LLaMA-1B	Standard RAG	0.4298†*	0.3032†*	0.2263	0.1064	0.2520*	0.2110	0.4083	0.2671	0.1839*	0.3726
	DA-RAG	0.3594†*	0.2587†*	0.2266	0.0869†*	0.2531*	0.1716*	0.3713†*	0.2221	0.2012*	0.3691
	FLARE	0.4013†*	0.2589†*	0.1960	0.0823†*	0.2234*	0.1630*	0.3784†*	0.1785*	0.1301†*	0.3173*
	DRAGIN	0.4556	0.3357*	0.1919	0.0901	0.2692*	0.1431*	0.4015	0.1830*	0.1056†*	0.3900
	P-RAG (Ours)	0.4920	0.3994	0.2185	0.1334	0.2764	0.1602*	**0.4493**	0.1999*	0.2205*	0.3482*
	Combine Both	**0.5046**	**0.4595**	**0.2399**	**0.1357**	**0.3237**	**0.2282**	0.4217	**0.2689**	**0.2961**	**0.4101**
Qwen-1.5B	Standard RAG	0.3875†*	0.3884†*	0.1187†*	0.0568†*	0.2431†*	0.1619*	0.3713†*	0.2073*	0.0999†*	0.2823*
	DA-RAG	0.3418†*	0.4015	0.1269†*	0.0514†*	0.2156†*	0.1182†*	0.3041†*	0.1683*	0.1197†*	0.2718†*
	FLARE	0.1896†*	0.1282†*	0.0852†*	0.0437†*	0.1004†*	0.0750†*	0.1229†*	0.0698†*	0.0641†*	0.1647†*
	DRAGIN	0.2771†*	0.1826†*	0.1025†*	0.0680†*	0.1538†*	0.0801†*	0.1851†*	0.0973†*	0.0548†*	0.1788†*
	P-RAG (Ours)	**0.4529**	**0.4494**	**0.2072**	**0.1372**	**0.3025**	0.1720	0.4623	0.2165*	0.1885	0.3280
	Combine Both	0.4053	0.4420	0.1705	0.1154	0.2627	**0.2383**	**0.5037**	**0.2942**	**0.2261**	**0.3495**
LLaMA-8B	Standard RAG	0.5843†*	0.4794†*	0.1833†*	0.0991†*	0.3372†*	0.1823†*	0.3493†*	0.2277†*	0.1613†*	0.3545†*
	DA-RAG	0.4921†*	0.3344†*	0.1523†*	0.0670†*	0.2396†*	0.1587†*	0.2860†*	0.1996†*	0.2255*	0.3481†*
	FLARE	0.4293†*	0.3769†*	0.3086	0.1627*	0.3492*	0.2493†*	0.4324†*	0.2771†*	0.2393*	0.3084†*
	DRAGIN	0.5185†*	0.4480†*	0.2664	0.1833	0.3544*	0.2618*	0.6116*	0.2924*	0.1772†*	0.3101†*
	P-RAG (Ours)	0.6353	0.5437	0.2471*	0.1992	0.3932	0.3115*	0.6557	0.3563*	0.2413*	0.4541
	Combine Both	**0.6432**	**0.5556**	**0.3160**	**0.2339**	**0.4258**	**0.4025**	**0.6918**	**0.4559**	**0.3059**	**0.4728**

实验效果

- 不同的lora权重初始化的策略对效果的影响

Table 2: Ablation study on the impact of LoRA weight initialization strategies for P-RAG. All metrics reported are F1 scores. "P-RAG Rand." and "P-RAG Warm." indicate randomly initialized LoRA weights and warm-up LoRA initialization, respectively. The best results are in bold.

		2WQA	HQA	PQA	CWQ
LLaMA-1B	**P-RAG Rand.**	0.2764	0.1999	**0.2205**	0.3482
	P-RAG Warm.	**0.3546**	**0.2456**	0.2035	**0.4263**
Qwen-1.5B	**P-RAG Rand.**	0.3025	0.2165	0.1885	0.3280
	P-RAG Warm.	**0.3542**	**0.2718**	**0.2418**	**0.5018**
LLaMA-8B	**P-RAG Rand.**	0.3932	0.3563	0.2413	0.4541
	P-RAG Warm.	**0.4201**	**0.4499**	**0.2952**	**0.5591**

- 随机参数初始化
- warm-up初始化：

用很小的学习率先简单训了一会，效果优于随机初始化

实验效果

- 文档增强阶段不同方法的消融实验

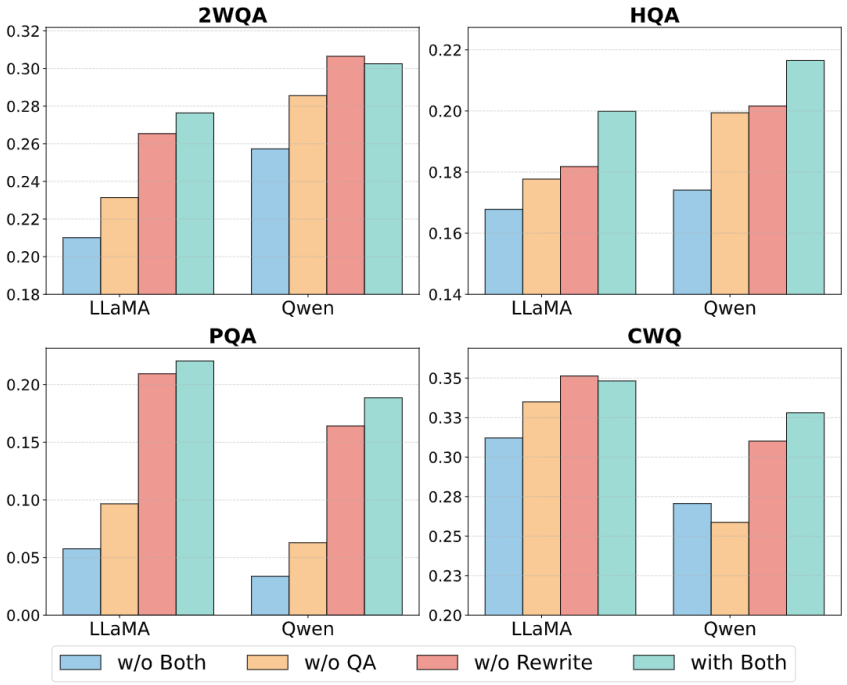


Figure 3: Ablation study on the impact of the document augmentation stage. LLaMA indicates LLaMA-3.2-1B, and Qwen indicates Qwen-2.5-1.5B. The metric used is the F1 Score.

Table 3: Ablation study comparing different document augmentation models. GenLM indicates the generator LLM and AugLM indicates the LLM for document augmentation. LLaMA indicates LLaMA-3.2-1B, and Qwen indicates Qwen-2.5-1.5B. The best results are in bold. The metric used in the table is F1 Score.

GenLM	AugLM	Dataset			
		2WQA	HQA	PQA	CWQ
LLaMA-1B	LLaMA-1B	0.2764	0.1999	0.2205	0.3482
	Qwen-1.5B	0.2753	0.1980	0.2340	0.3495
	LLaMA-8B	0.2748	0.1935	0.2207	0.3498
Qwen-1.5B	LLaMA-1B	0.2974	0.2005	0.1829	0.3183
	Qwen-1.5B	0.3025	0.2165	0.1885	0.3280
	LLaMA-8B	0.2948	0.2161	0.2156	0.3211

实验效果

- 加载参数这一步的计算量还不到前向传播一个token所需要的计算量

Table 4: The average time required by the LLaMA3-8B model to answer a question on the 2WikiMultihopQA (2WQA) and ComplexWebQuestions (CWQ) datasets. The "+0.32" footnote for P-RAG and Combine Both indicates the total time needed for merging and loading the LoRA adapter.

	2WQA		CWQ	
	Time(s)	**Speed Up**	**Time(s)**	**Speed Up**
P-RAG	$2.34_{+0.32}$	1.29x	$2.07_{+0.32}$	1.36x
Combine Both	$3.08_{+0.32}$	0.98x	$2.84_{+0.32}$	0.99x
Standard RAG	3.03	1.00x	2.82	1.00x
FLARE	10.14	0.25x	11.31	0.25x
DRAGIN	14.60	0.21x	16.21	0.17x

总结

- 相当于实现了一个动态LLM：
  - 不同的query对应不同的LLM参数
- 提升效率：
  - PRAG不需要占用任何context
  - 文档参数化可以离线做，且plug-in的成本不高
- 通过lora技术将外部知识高效编码到模型参数中，解决了传统RAG上下文冗余问题

Thanks !