Parametric Retrieval Augmented Generation

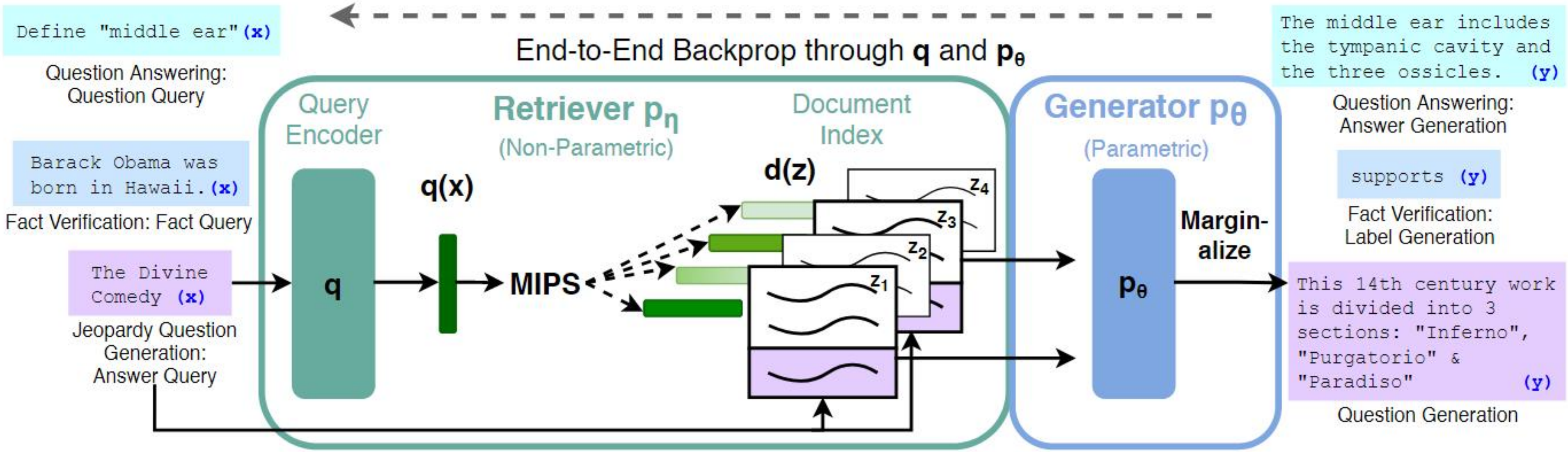
2025.3.26

唐明昊

Su, Weihang, et al. "Parametric Retrieval Augmented Generation." arXiv preprint arXiv:2501.15915 (2025).

Motivation

All RAG methods, regardless of their variations, share a common characteristic: they inject external knowledge by directly adding passages or documents into the input context of LLMs, which we refer to as the **in-context knowledge injection**


Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and p_θ
Query Encoder
q
q(x)
Retriever p_η
(Non-Parametric)
MIPS
Document Index
d(z)
z_4
z_3
z_2
z_1
Generator p_θ
(Parametric)
p_θ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Lewis, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.

Motivation

- Long context not only introduces extra **computational overhead** and **latency** for LLM inference, but also **hurts the performance** of LLMs in understanding and utilizing external knowledge

- Adding passages in the input context could **only affect the online computation of key-value** pairs in the attention networks of LLMs, but not the model's stored parameters, where its knowledge is encoded

Motivation

LLMs may never be able to utilize external knowledge as effectively as they use their internal knowledge in in–context RAG methods

SFT–based methods?

require substantial resources, may affect original ability, lack of flexibility

Is it possible to inject external knowledge into LLM parameters effectively, efficiently, and flexibly for retrieval–augmented generation?

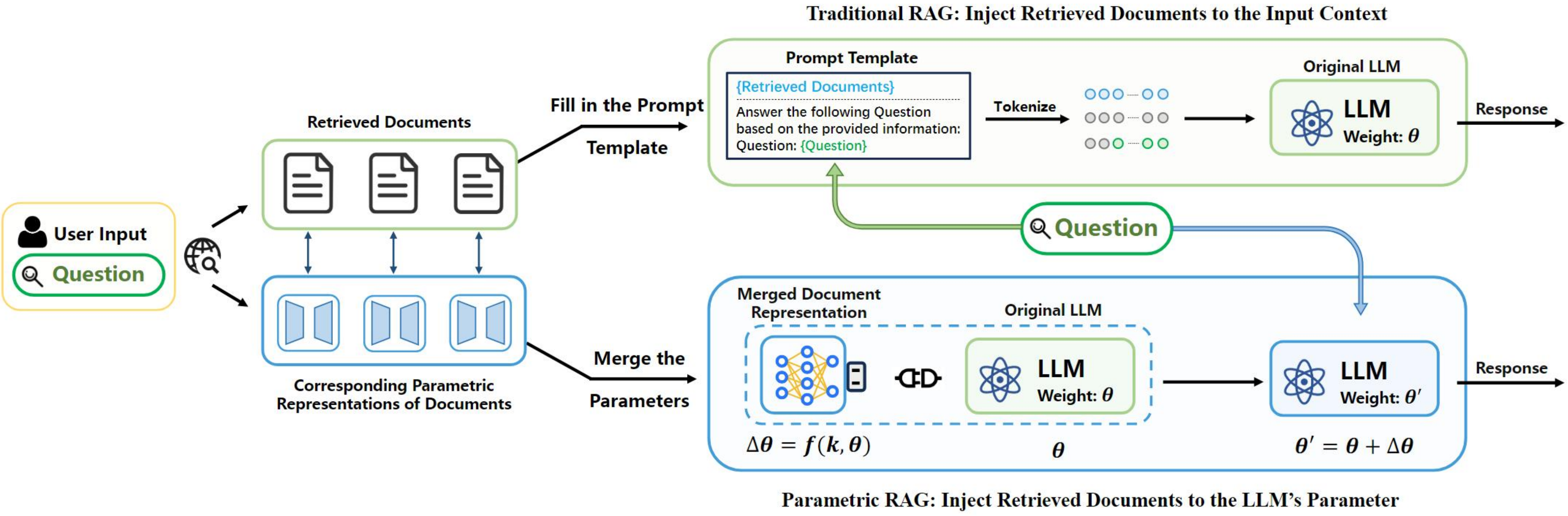
Overview

- offline document parameterization

  converts each document into its corresponding parametric representation

- online inference with a Retrieve–Update–Generate workflow

  merges the parametric representations and then plugs the merged parameters into the LLM

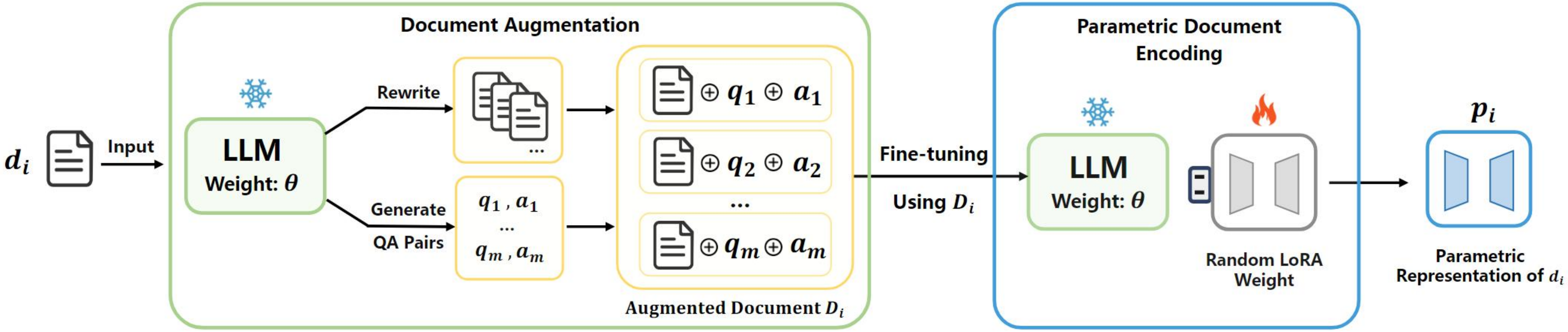

Traditional RAG: Inject Retrieved Documents to the Input Context
Prompt Template
{Retrieved Documents}
Answer the following Question based on the provided information:
Question: {Question}
Tokenize
Original LLM
LLM
Weight: θ
Response
User Input
Question
Retrieved Documents
Fill in the Prompt
Template
Corresponding Parametric Representations of Documents
Merge the
Parameters
Question
Merged Document Representation
Original LLM
LLM
Weight: θ
LLM
Weight: θ'
Response
$\Delta\theta = f(k, \theta)$
θ
$\theta' = \theta + \Delta\theta$
Parametric RAG: Inject Retrieved Documents to the LLM's Parameter

Offline Document Parameterization

Document Augmentation

- Document Rewriting
- QA Pair Generation

$$D_i = \{(d_i{}^k, q_i{}^j, a_i{}^j) \mid 1 \le k \le n,\ 1 \le j \le m\}$$

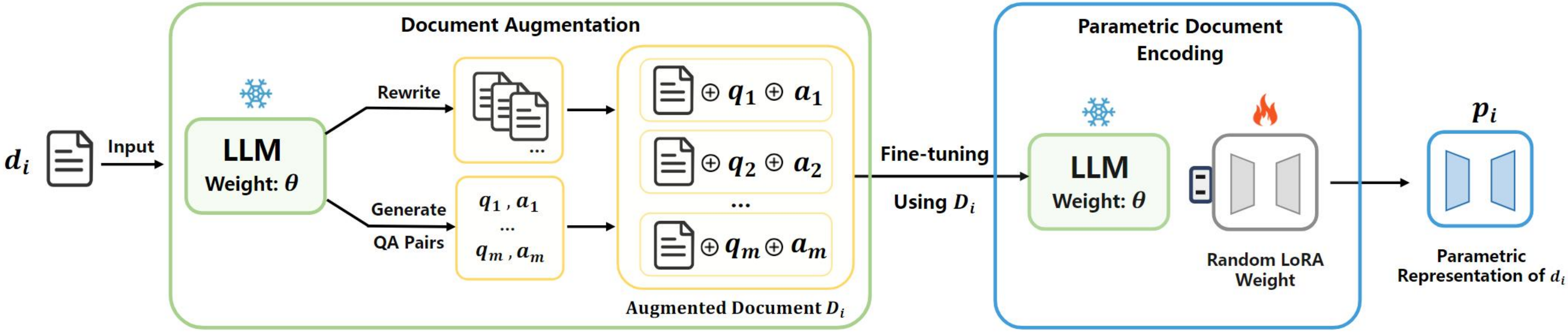


Offline Document Parameterization

Parametric Document Encoding

- initialize parametric representations as low-rank matrices corresponding to FFN, following the LoRA approach
- each document is associated with independently trained low-rank parameters

$$W' = W + \Delta W = W + A B^{\top} \qquad \min_{\Delta\theta} \sum_{(d_i{}^k, q_i{}^j, a_i{}^j) \in D_i} \sum_{t=1}^{T} -\log P_{\theta+\Delta\theta}(x_t \mid x_{<t})$$



Online Inference

- Retrieve

  select the top-k documents with the highest relevance scores

- Update

  merge the low-rank matrices from the top-k retrieved documents to form a single plug-in module for the LLM, and then update the original feed-forward weight

$$\Delta W_{\text{merge}} = \alpha \cdot \sum_{j=1}^{k} A_j \, B_j^{\top} .$$

- Generate

  directly generate the final response to the query

Experimental Results

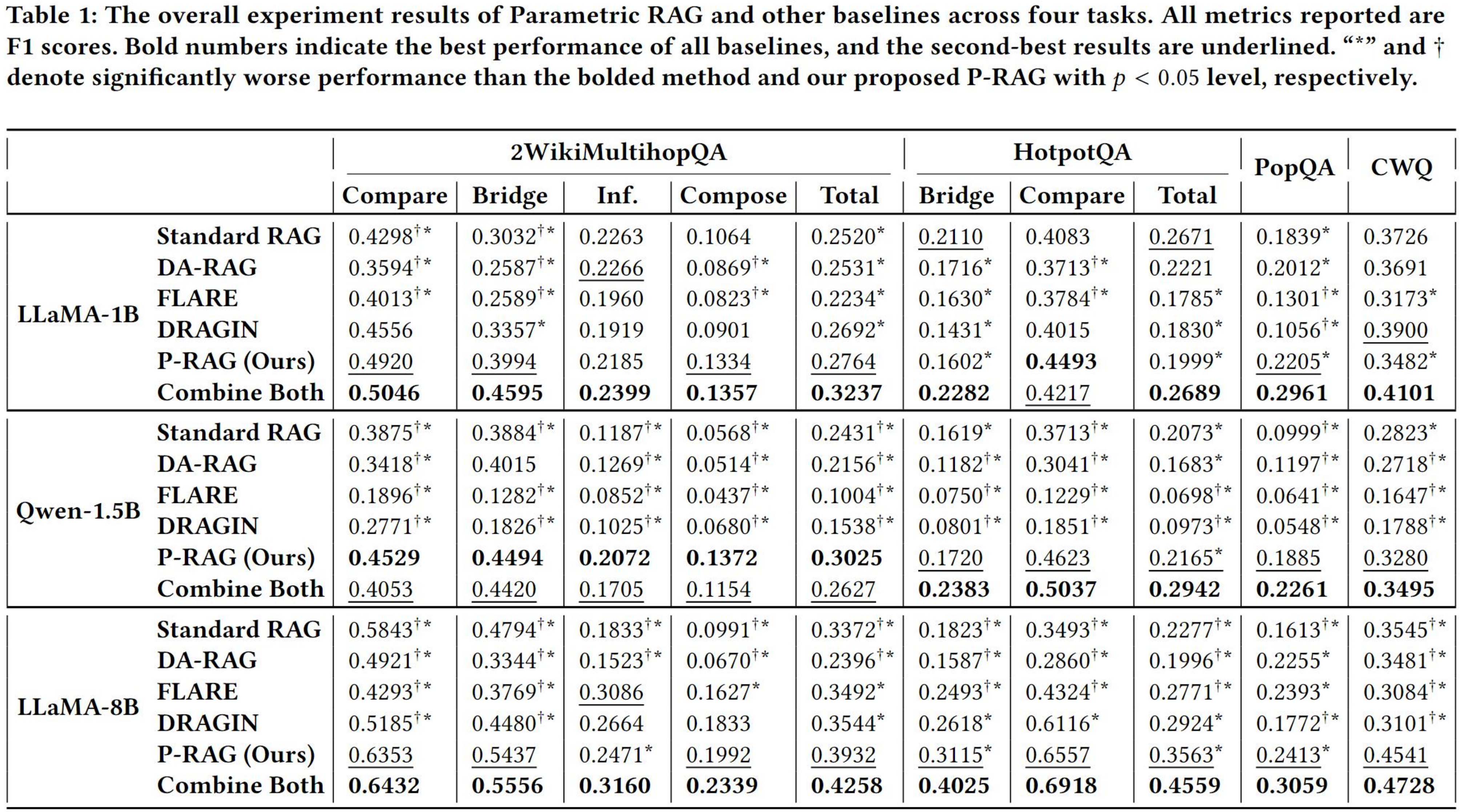


Table 1: The overall experiment results of Parametric RAG and other baselines across four tasks. All metrics reported are F1 scores. Bold numbers indicate the best performance of all baselines, and the second-best results are underlined. “*” and † denote significantly worse performance than the bolded 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	$0.4298^{\dagger *}$	$0.3032^{\dagger *}$	0.2263	0.1064	0.2520^{*}	0.2110	0.4083	0.2671	0.1839^{*}	0.3726
	DA-RAG	$0.3594^{\dagger *}$	$0.2587^{\dagger *}$	0.2266	$0.0869^{\dagger *}$	0.2531^{*}	0.1716^{*}	$0.3713^{\dagger *}$	0.2221	0.2012^{*}	0.3691
	FLARE	$0.4013^{\dagger *}$	$0.2589^{\dagger *}$	0.1960	$0.0823^{\dagger *}$	0.2234^{*}	0.1630^{*}	$0.3784^{\dagger *}$	0.1785^{*}	$0.1301^{\dagger *}$	0.3173^{*}
	DRAGIN	0.4556	0.3357^{*}	0.1919	0.0901	0.2692^{*}	0.1431^{*}	0.4015	0.1830^{*}	$0.1056^{\dagger *}$	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	$0.3875^{\dagger *}$	$0.3884^{\dagger *}$	$0.1187^{\dagger *}$	$0.0568^{\dagger *}$	$0.2431^{\dagger *}$	0.1619^{*}	$0.3713^{\dagger *}$	0.2073^{*}	$0.0999^{\dagger *}$	0.2823^{*}
	DA-RAG	$0.3418^{\dagger *}$	0.4015	$0.1269^{\dagger *}$	$0.0514^{\dagger *}$	$0.2156^{\dagger *}$	$0.1182^{\dagger *}$	$0.3041^{\dagger *}$	0.1683^{*}	$0.1197^{\dagger *}$	$0.2718^{\dagger *}$
	FLARE	$0.1896^{\dagger *}$	$0.1282^{\dagger *}$	$0.0852^{\dagger *}$	$0.0437^{\dagger *}$	$0.1004^{\dagger *}$	$0.0750^{\dagger *}$	$0.1229^{\dagger *}$	$0.0698^{\dagger *}$	$0.0641^{\dagger *}$	$0.1647^{\dagger *}$
	DRAGIN	$0.2771^{\dagger *}$	$0.1826^{\dagger *}$	$0.1025^{\dagger *}$	$0.0680^{\dagger *}$	$0.1538^{\dagger *}$	$0.0801^{\dagger *}$	$0.1851^{\dagger *}$	$0.0973^{\dagger *}$	$0.0548^{\dagger *}$	$0.1788^{\dagger *}$
	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	$0.5843^{\dagger *}$	$0.4794^{\dagger *}$	$0.1833^{\dagger *}$	$0.0991^{\dagger *}$	$0.3372^{\dagger *}$	$0.1823^{\dagger *}$	$0.3493^{\dagger *}$	$0.2277^{\dagger *}$	$0.1613^{\dagger *}$	$0.3545^{\dagger *}$
	DA-RAG	$0.4921^{\dagger *}$	$0.3344^{\dagger *}$	$0.1523^{\dagger *}$	$0.0670^{\dagger *}$	$0.2396^{\dagger *}$	$0.1587^{\dagger *}$	$0.2860^{\dagger *}$	$0.1996^{\dagger *}$	0.2255^{*}	$0.3481^{\dagger *}$
	FLARE	$0.4293^{\dagger *}$	$0.3769^{\dagger *}$	0.3086	0.1627^{*}	0.3492^{*}	$0.2493^{\dagger *}$	$0.4324^{\dagger *}$	$0.2771^{\dagger *}$	0.2393^{*}	$0.3084^{\dagger *}$
	DRAGIN	$0.5185^{\dagger *}$	$0.4480^{\dagger *}$	0.2664	0.1833	0.3544^{*}	0.2618^{*}	0.6116^{*}	0.2924^{*}	$0.1772^{\dagger *}$	$0.3101^{\dagger *}$
	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**
]

Experimental Results

LoRA initialization: random or pretrained using sampled QA pairs

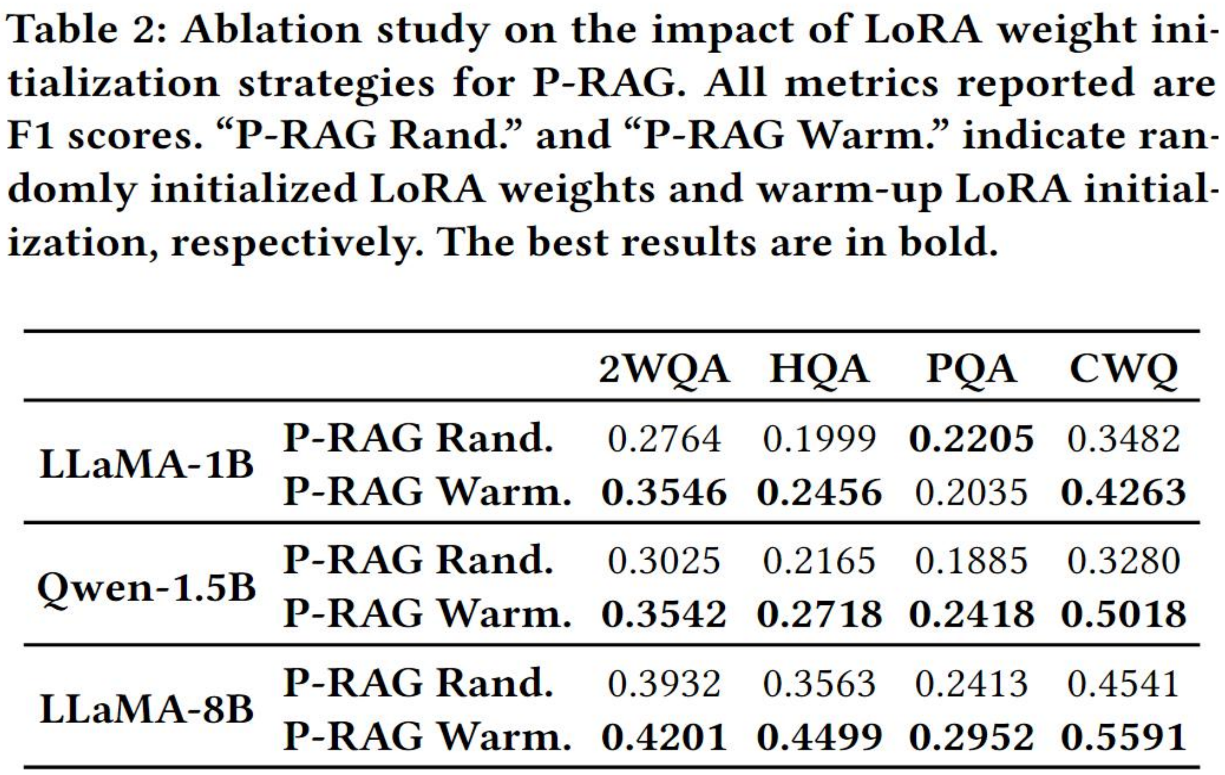


Table 2: Ablation study on the impact of LoRA weight initialization strategies for P-RAG. All metrics reported are F1 scores. “P-RAG Rand.” and “P-RAG Warm.” indicate randomly initialized LoRA weights and warm-up LoRA initialization, respectively. The best results are in bold.

		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**

Experimental Results

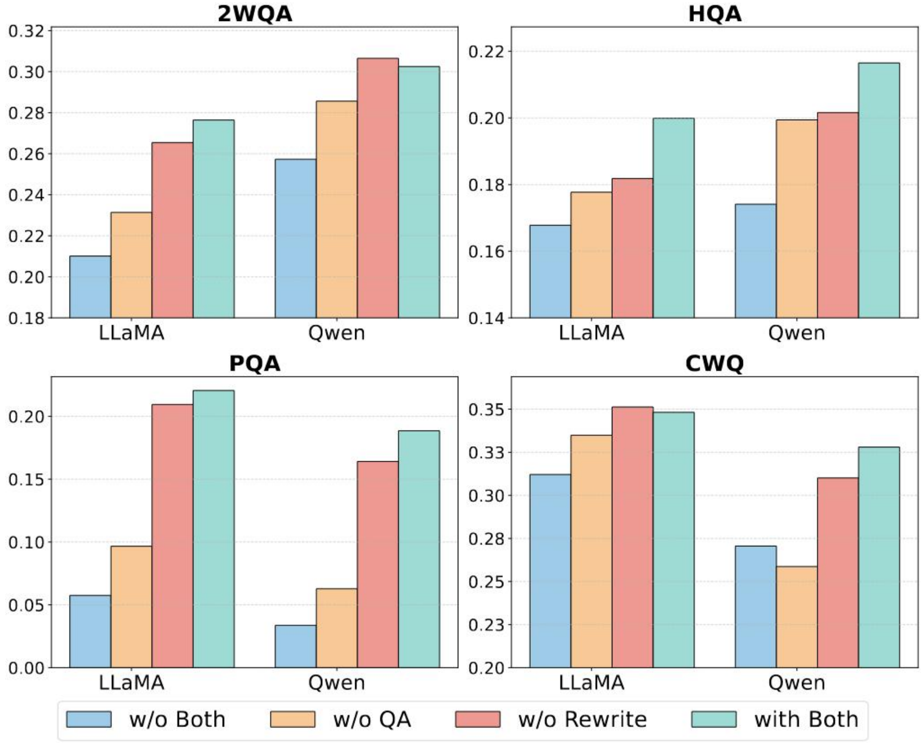


Figure 3: Ablation study on the impact of the document augmentation stage. LLaMA indicates LLaMA-3.2-1B, and Qwen indicates Qwen-2.5-1.5B. The metric used is the F1 Score.

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

Efficiency

- Parametric RAG is more cost–friendly than in–context RAG when the number of queries is more than twice that of documents in the life cycle of the service
- The access of information in real user traffic follows a longtail distribution, creating parametric representations for a tiny set of head documents can serve the majority of user requests

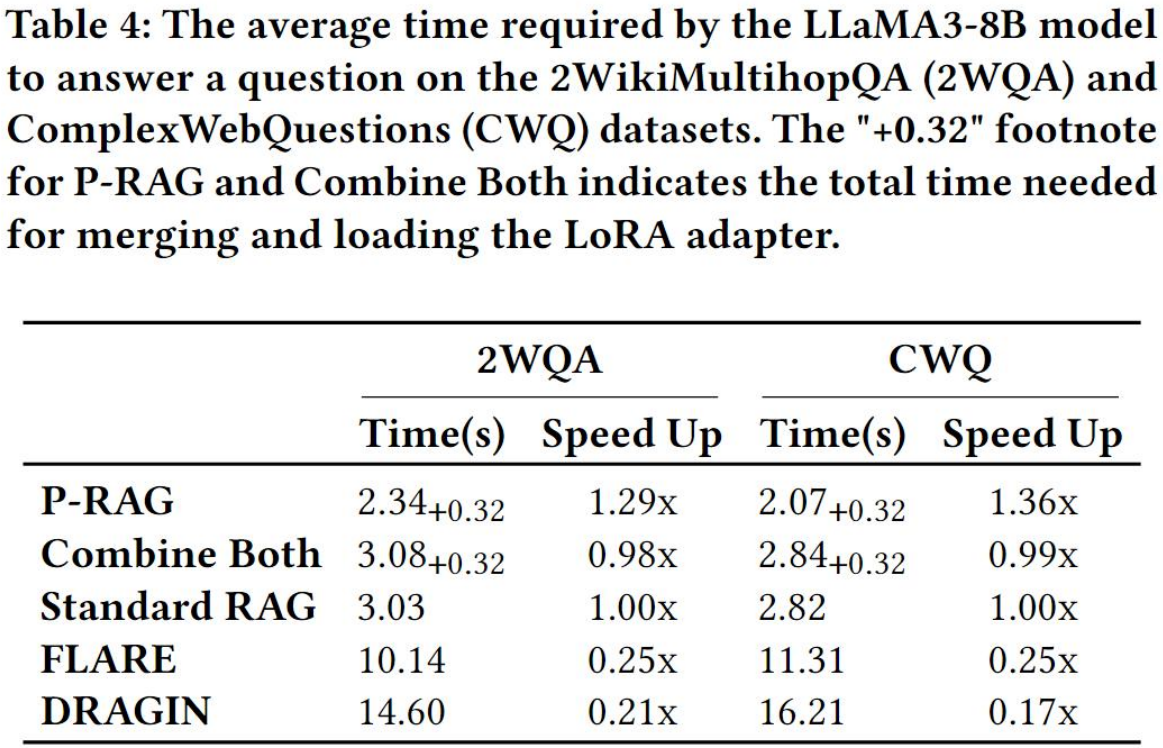


Table 4: The average time required by the LLaMA3-8B model to answer a question on the 2WikiMultihopQA (2WQA) and ComplexWebQuestions (CWQ) datasets. The "+0.32" footnote for P-RAG and Combine Both indicates the total time needed for merging and loading the LoRA adapter.

	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