

LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering

Qingfei Zhao^{1,2,†}, Ruobing Wang^{1,2}, Yukuo Cen⁴,
Daren Zha¹, Shicheng Tan³, Yuxiao Dong³, Jie Tang^{3,*}

¹Institute of Information Engineering, Chinese Academy of Sciences;

²School of Cyber Security, University of Chinese Academy of Sciences;

³Tsinghua University; ⁴Zhipu AI

{zhaoqingfei, wangruobing, zhadaren}@iie.ac.cn, yukuo.cen@zhipuai.cn

tsctan@foxmail.com, {yuxiaod, jietang}@tsinghua.edu.cn

EMNLP 2024

HUMANE Lab

김태균


2025.01.17



Background

- Existing long-context LLMs for LCQA often struggle with the “lost in the middle” issue
- RAG mitigates this issue by employing a “fixed-length chunking strategy”
- However, its chunking strategy in long-context disrupts the “global information”, and its low-quality retrieval hinders LLMs from identifying effective “factual details” due to substantial noise

Background

 **Question:** Where did the performer of song I' ll Say It graduate from? ➔ **Thought:** I' ll Say It → Griffin → Lee Strasberg Theatre and Film Institute
Answer: Lee Strasberg Theatre and Film Institute

LongRAG



Integrated Information: I' ll Say It is a song... recorded by comedian Kathy Griffin. ... She performer of the song "I' ll Say It" is Kathy Griffin. She attended the Lee Strasberg Theatre and Film Institute in Los Angeles, where she studied drama. ✓

Answer: Lee Strasberg Theatre and Film Institute

Vanilla RAG



Retrieved Information: I' ll Say It is a song... recorded by comedian Kathy Griffin. ... She became an adjunct professor and part-time lecturer at Seoul Arts College. ✗

Incomplete Key Information

Answer: Seoul Arts College

Long-Context QA



Long-Context Information: I' ll Say It is a song... recorded by comedian Kathy Griffin. ... Griffin ... studied drama at the Lee Strasberg Theatre and Film Institute. ... (too long context) ... Song Yoon-ah ... as a freshman at Hanyang University. ✗

Answer: Hanyang University

Lost In the Middle

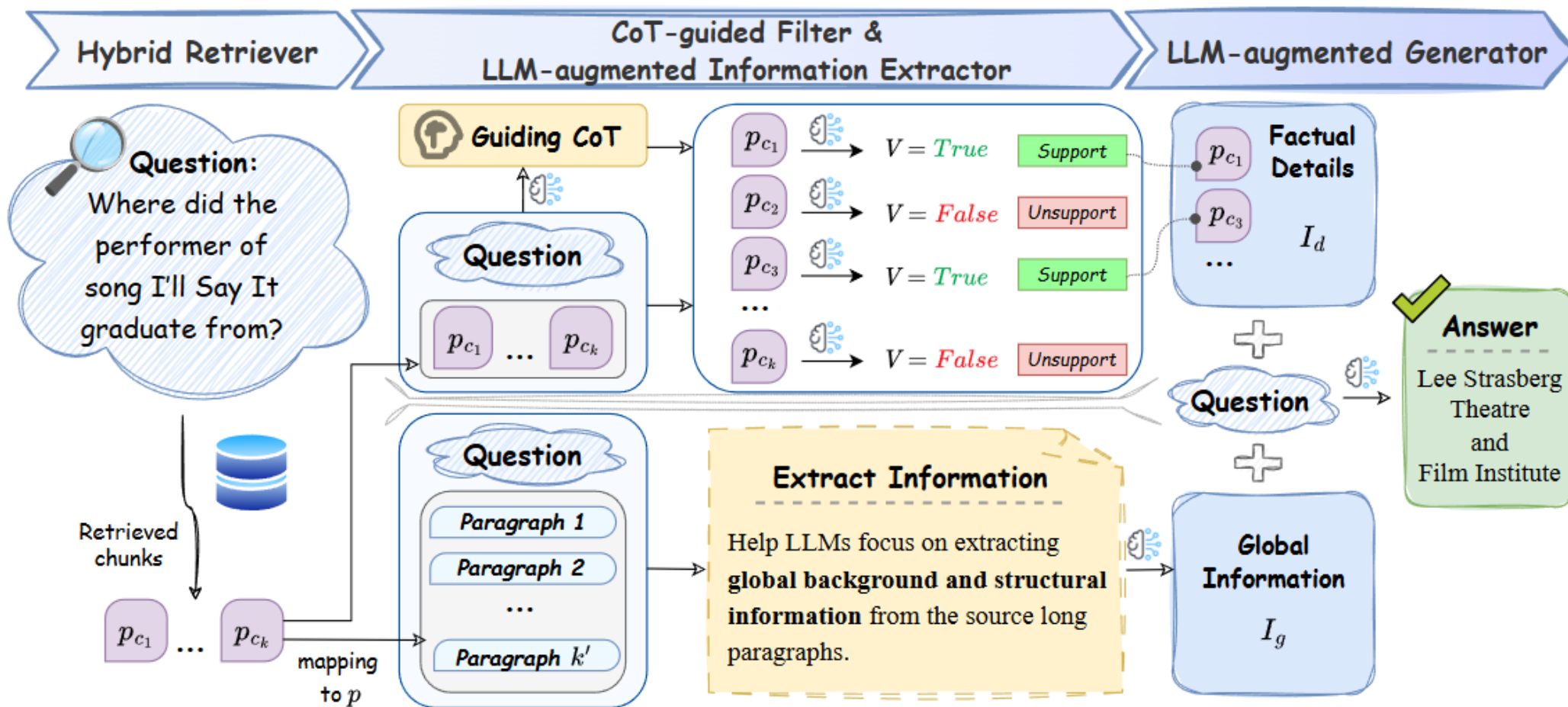
produces an incorrect response

Background

- Several advanced RAG systems have attempted to mitigate the issues
 - Self-RAG
 - CRAG
- However, challenges remain as they may still filter out relevant chunks or miss factual details

⇒ LongRAG

Overview



LRGInstruction

Instruction-following dataset for fine-tuning

- Construction pipeline is automated
 - Facilitating adaptation(transferability) to various domains and LLMs

LRGInstruction

- Training sets from multi-hop datasets
 - HotpotQA
 - 2WikiMultiHopQA
 - MusiQue
 - QASPER

Datasets	HotpotQA	2WikiMultiHopQA	MusiQue	QASPER
Num of long-context extractor data	200	200	200	100
Num of CoT-guiding data	200	200	200	100
Num of filtering data	200	200	200	-
Num of task-oriented data	200	200	200	-
Num of samples	800	800	800	200

Table 15: Statistics of our fine-tuning instruction dataset LRGInstruction.

LRGInstruction

Types of data

1. Long-context extractor data
2. CoT-guiding & Filtering data
3. Task-oriented data

LRGInstruction

1. Long-context extractor data

- p_s : supporting paragraphs
- I_g : global information
- [STEP] : prompt
- [RESULT] : gold data

[STEP-1]: Data construction prompt for Extractor
<code>{supporting paragraphs}</code>
Based on the above background only, please output the original information that needs to be cited to answer the following questions. Please ensure that the information cited is detailed and comprehensive.
Question: <code>{question}</code>
Output only the original information of the required reference: <code>{global information}</code>
[STEP-2]: An LLM-based self-evaluator for Extractor
I am going to provide you with a question, the background information, and the answer to that question. Please evaluate whether the answer can be solely derived from the given background information. If it can, set the status value as True, if it can't, set the status value as False.
Question: <code>{question}</code>
Background Information: <code>{global information}</code>
Answer: <code>{answer}</code>
Your output format should be the following json format: status: {the value of status}

[RESULT]: Long-Context Extractor Data for Extractor
Instruction: <code>{content}</code> Based on the above background, please output the information you need to cite to answer the question below. <code>{question}</code>
Output: <code>{global information}</code>

Table 16: Data construction pipeline for extractor and format illustration of long-context extractor data.

LRGInstruction

2. CoT-guiding & Filtering data

<p>[STEP-1]: Data construction prompt for CoT guidance stage</p> <p>{supporting paragraphs}</p> <p>Given question:{question}</p> <p>The answer is:{answer}</p> <p>Your task is to give your thought process for this given question based on the above information, only give me your thought process and do not output other information.</p> <p>Thought process: {CoT}</p>
<p>[STEP-2]: An LLM-based self-evaluator for CoT guidance stage</p> <p>Question:{question}</p> <p>Thought process of the question:{CoT}</p> <p>Answer:{answer}</p> <p>Please evaluate whether the thought process of this question can explain the answer to this question. If it can explain the answer, set the value of status to True. If it cannot explain the answer, set the value of status to False. Your output format should be the following json format:</p> <p>status: {the value of status}</p>

<p>[RESULT-1]: CoT-guiding Data for CoT guidance stage</p> <p>Instruction:</p> <p>{content}</p> <p>Please combine the above information and give your thought process for the following</p> <p>Question:{question}</p> <p>Output:</p> <p>{CoT}</p>
<p>[RESULT-2]: Filtering Data for filtering stage</p> <p>Instruction:</p> <p>Given an article:{content}</p> <p>Question:{question}</p> <p>Thought process for the question:{CoT}</p> <p>Your task is to use the thought process provided to decide whether you need to cite the article to answer this question. If you need to cite the article, set the status value to True. If not, set the status value to False. Please output the response in the following json format:</p> <p>{"status": {the value of status}}</p> <p>Output:</p> <p>{status}</p>

Table 17: Data construction pipeline for filter, and format illustration of CoT-guiding and filtering data.

LRGInstruction

3. Task-oriented data

[RESULT]: Task-Oriented Data for RAG task	
Instruction:	
{content}	
Based on the above information, Only give me the answer and do not output any other words.	
Question:	{question}
Output:	
{answer}	

Table 18: Data construction pipeline for RAG task, and format illustration of task-oriented data.

LongRAG

A general, dual-perspective, and robust LLM-based RAG system for LCQA

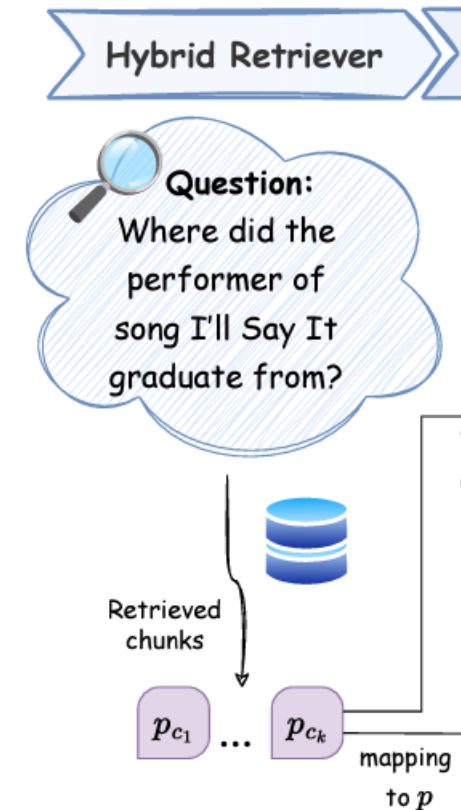
1. Hybrid retriever
2. LLM-augmented information extractor
3. CoT-guided filter
4. LLM-augmented generator

LongRAG

1. Hybrid retriever

: Given question, recalls k chunks (p_c)

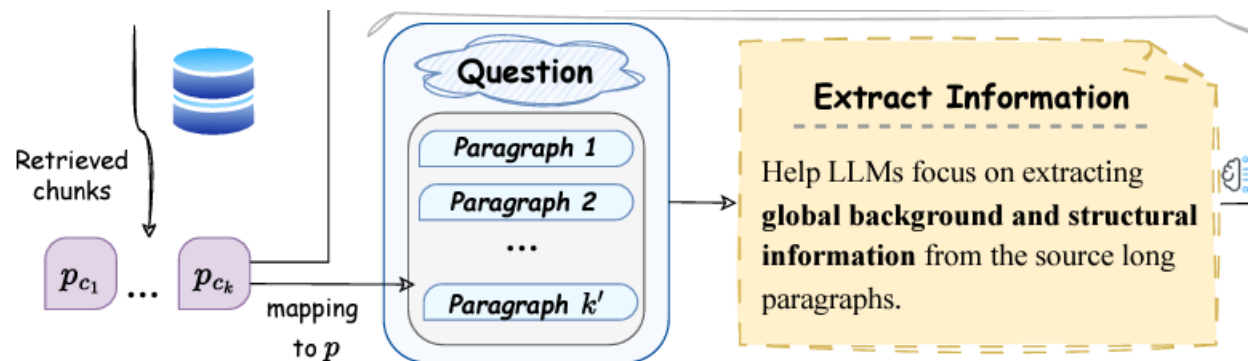
- Dual-encoder : For rapid retrieval at a coarse-grained level
- Cross-encoder : To capture the deep semantic interaction



LongRAG

2. LLM-augmented information extractor

- Map the chunks (p_c) to their source long-context paragraphs (p)
 - $f_m(p_{c_1}, p_{c_2}, \dots, p_{c_k}) \rightarrow p_1, p_2, \dots, p_{k'}$
- Then concatenate k' paragraphs and feed them into prompt of the LLM-augmented information extractor
 - $I_g = LLM(prompt_c(q, p_1 || p_2 || \dots || p_{k'}))$



LongRAG

3. CoT-guided filter

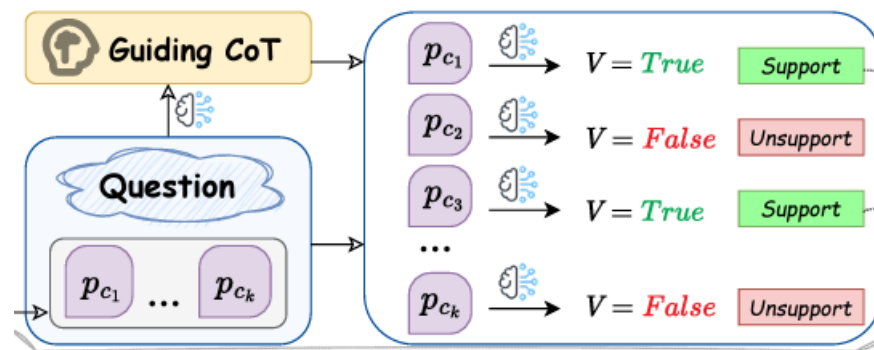
- CoT guidance : Generates a CoT with a global perspective based on the retriever semantic space p_c

- $CoT = LLM(prompt_c(q, p_{c_1} || p_{c_2} || \dots || p_{c_k}))$

- Filtering stage : The relevance between the question and the chunks is evaluated to remove unnecessary chunks and retain factual details

- $V(q, p_c, CoT) = \begin{cases} \text{True}, & \text{if support} \\ \text{False}, & \text{otherwise} \end{cases}$

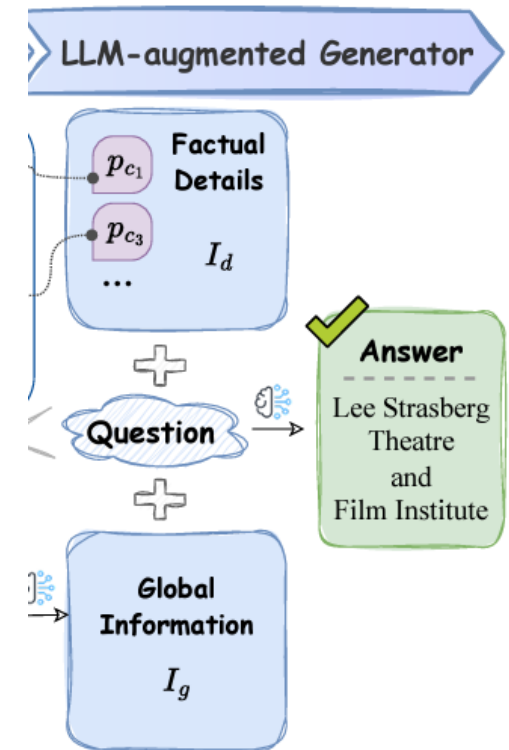
- $I_d = \{p_c \mid V(q, p_c, CoT) = \text{True}\}$



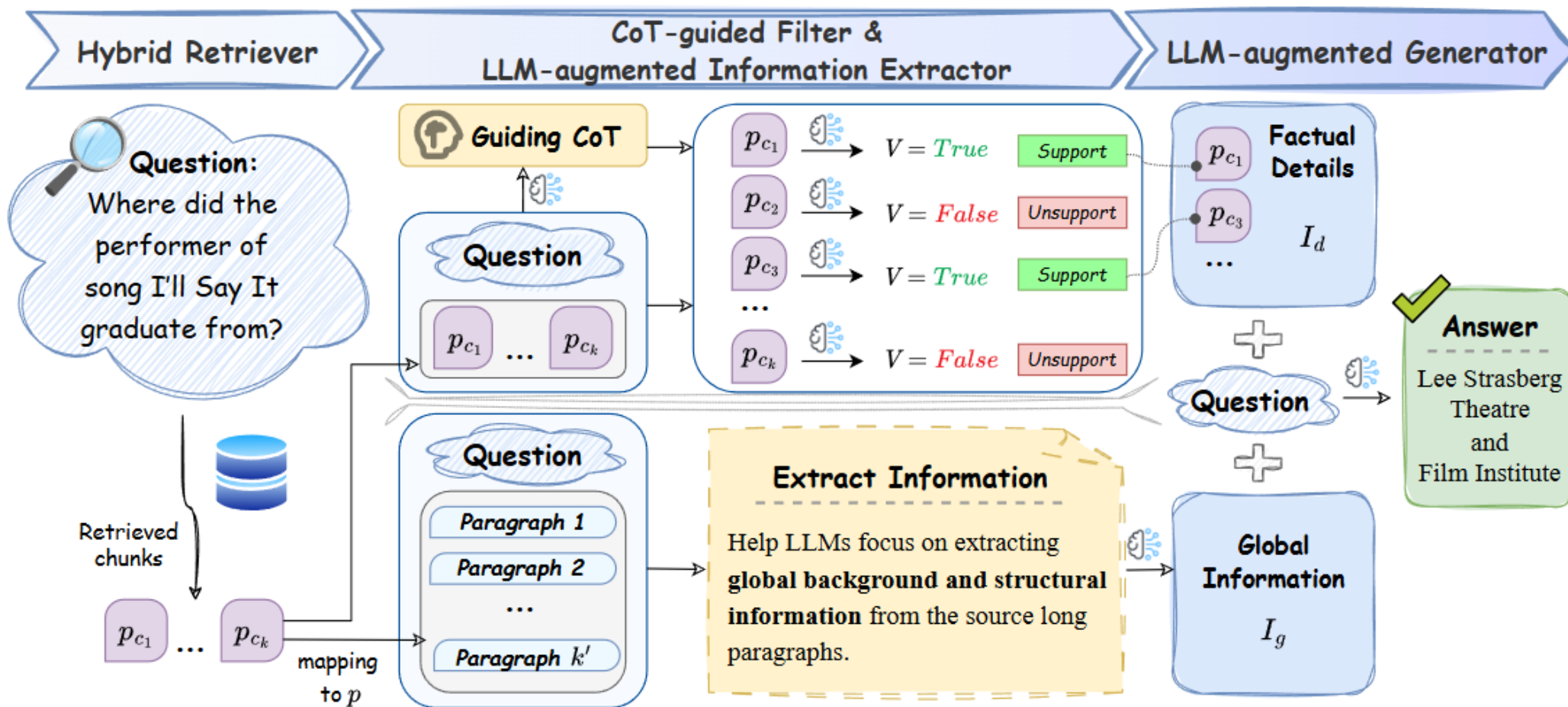
LongRAG

4. LLM-augmented generator

- The generator boosts the interaction of knowledge across these two perspectives (I_g , I_d) to produce answers to questions
 - $\alpha = LLM(prompt_g(I_g, I_d))$



LongRAG



Experiment

Experimental setup

- Evaluation metric : F1-score
- Categories of baselines
 1. Long-context LLM methods
 2. Advanced RAG methods
 3. Vanilla RAG

Experiment

Overall performance

1. Long-context LLM
2. Other RAG
3. Model size

Model	HotpotQA	2WikiMQA	MusiQue	Average
# Long-Context LLM Methods #				
LongAlign-7B-64k (<i>Llama2</i>) (Bai et al., 2024)	48.85	28.56	25.14	34.18
LongLoRA-13B-32k (<i>Llama2</i>) (Chen et al., 2023b)	47.45	42.92	29.46	39.94
# Advanced RAG Methods #				
CFIC-7B (<i>Llama2</i>) (Qian et al., 2024)	34.00	-	14.70	24.35
CRAG (<i>GPT-3.5-Turbo</i>) (Yan et al., 2024)	52.04	41.13	25.34	39.50
Self-RAG (<i>GPT-3.5-Turbo</i>) (Asai et al., 2023)	50.51	46.75	24.62	40.63
# RAG-Base (Vanilla RAG) #				
Vicuna-v1.5-7B-16k (Zheng et al., 2023)	38.63	27.92	15.68	27.41
Qwen-1.5-7B-32k (Bai et al., 2023a)	45.70	34.69	25.08	35.16
Llama3-8B-8k (Touvron et al., 2023)	48.25	43.47	19.66	37.13
ChatGLM3-6B-32k (Du et al., 2022)	52.57	42.56	25.51	40.21
GPT-3.5-Turbo-16k	50.17	45.32	21.84	39.11
GPT-3.5-Turbo	52.31	43.44	25.22	40.32
Llama3-70B-8k	52.33	50.23	25.49	42.68
GLM-4	57.41	52.91	27.55	45.96
# Ours with SFT #				
LongRAG-Llama2-7B-4k	53.85	45.61	26.22	41.89
LongRAG-Llama2-13B-4k	57.05	49.95	33.63	46.88
LongRAG-Qwen-1.5-7B-32k	52.91 (7.21↑)	46.65 (11.96↑)	31.85 (6.77↑)	43.80 (8.65↑)
LongRAG-Llama3-8B-8k	52.39 (4.14↑)	49.67 (6.20↑)	31.70 (12.04↑)	44.59 (7.46↑)
LongRAG-Vicuna-v1.5-7B-16k	55.55 (16.92↑)	50.13 (22.21↑)	28.29 (12.61↑)	44.66 (17.25↑)
LongRAG-ChatGLM3-6B-32k	55.93 (3.36↑)	54.85 (12.29↑)	33.00 (7.49↑)	47.93 (7.71↑)
# Ours without SFT #				
LongRAG-GPT-3.5-Turbo	56.17 (3.86↑)	51.37 (7.93↑)	32.83 (7.61↑)	46.79 (6.47↑)
LongRAG-GPT-3.5-Turbo-16k	59.11 (8.94↑)	51.25 (5.93↑)	30.37 (8.53↑)	46.91 (7.80↑)
LongRAG-GLM-4	62.11 (4.70↑)	57.16 (4.25↑)	38.40 (10.85↑)	52.56 (6.60↑)

Experiment

Ablation study

1. R&L
2. Ext. vs R&L
3. E&F vs Others

Model	HotpotQA					2WikiMQA					MusiQue				
	R&B	R&L	Ext.	Fil.	E&F	R&B	R&L	Ext.	Fil.	E&F	R&B	R&L	Ext.	Fil.	E&F
# Ours with SFT #															
LongRAG-ChatGLM3-6B-32k	51.48	54.00	<u>55.11</u>	49.01	55.93	46.61	44.83	<u>52.53</u>	48.83	54.85	24.02	33.15	32.98	27.70	<u>33.00</u>
LongRAG-Qwen1.5-7B-32k	47.09	48.93	<u>50.01</u>	49.11	52.91	35.78	37.72	<u>42.91</u>	38.98	46.65	20.68	26.08	<u>29.60</u>	23.67	31.85
LongRAG-Vicuna-v1.5-7B-16k	51.63	50.18	55.94	52.34	<u>55.55</u>	39.45	43.53	<u>49.57</u>	41.18	50.13	25.30	25.28	<u>29.25</u>	29.29	28.29
LongRAG-Llama3-8B-8k	49.45	50.49	<u>51.77</u>	49.64	52.39	39.79	37.16	<u>46.80</u>	42.40	49.67	21.41	22.90	33.85	23.47	<u>31.70</u>
# Ours without SFT #															
LongRAG-ChatGLM3-6B-32k	<u>52.57</u>	50.19	52.27	53.36	52.07	42.56	42.92	<u>44.95</u>	42.94	46.08	25.51	29.93	28.27	23.99	<u>28.45</u>
LongRAG-Qwen1.5-7B-32k	45.70	49.72	<u>50.74</u>	45.70	50.80	34.69	<u>35.49</u>	39.53	34.69	39.53	<u>25.08</u>	25.85	29.75	<u>25.08</u>	29.75
LongRAG-Vicuna-v1.5-7B-16k	38.63	30.40	<u>41.45</u>	39.46	43.18	27.92	20.68	<u>29.08</u>	29.89	30.85	15.68	8.92	17.65	<u>16.35</u>	16.98
LongRAG-Llama3-8B-8k	48.25	48.72	52.44	47.75	<u>52.19</u>	43.47	41.59	47.34	42.22	<u>46.57</u>	19.66	23.62	<u>24.90</u>	20.06	24.99
LongRAG-GPT-3.5-Turbo	52.31	55.30	<u>56.15</u>	50.90	56.17	43.44	45.03	53.29	39.49	<u>51.37</u>	25.22	28.65	<u>32.17</u>	24.41	32.83
LongRAG-GPT-3.5-Turbo-16k	50.17	49.80	60.06	47.10	<u>59.11</u>	45.32	46.80	51.26	46.38	<u>51.25</u>	21.84	25.09	<u>26.92</u>	22.02	30.37
LongRAG-GLM-4	57.41	56.17	<u>61.07</u>	55.41	62.11	52.91	48.98	<u>54.22</u>	52.61	57.16	27.55	27.85	38.54	28.12	<u>38.40</u>

Experiment

Transferability

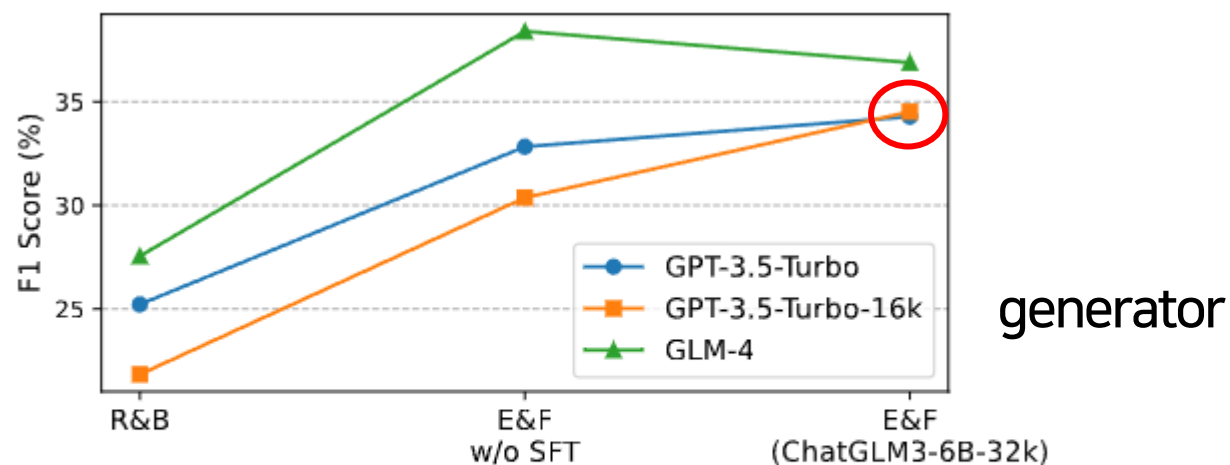


Figure 4: Analysis of the transferability of Extractor&Filter on dataset MusiQue.

Conclusion

LongRAG which enhances RAG's performance in LCQA tasks

- Addresses two main issues
 - The incomplete collection of long-context information
 - The difficulty in precisely identifying factual information

Open question

Will it also demonstrate effective performance for long-form answers?