LiR^{3}AG: A Lightweight Rerank Reasoning Strategy Framework for Retrieval-Augmented Generation

Guo Chen[1], Junjie Huang[1, *], Huaijin Xie[2], Fei Sun[3], Tao Jia[1, 4]

[1]College of Computer and Information Science, Southwest University, Chongqing, China
[2]School of Computer and Cyber Sciences, Communication University of China, Beijing, China
[3]State Key Laboratory of AI Safety, Institute of Computing Technology, CAS, Beijing, China
[4]College of Computer and Information Science, Chongqing Normal University
[*] Junjie Huang is the corresponding author

Memes

Introduction

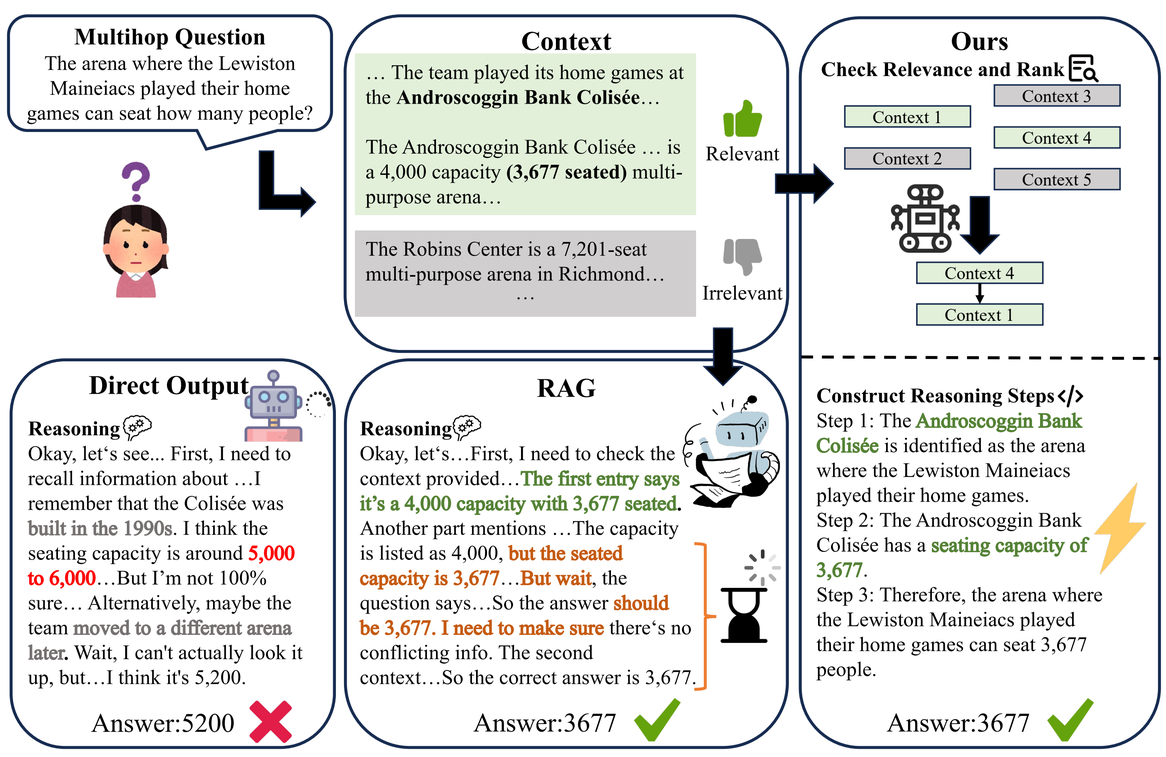

Figure. 1. A multi-hop QA example where direct generation hallucinates, RAG answers correctly but with redundant reasoning, and LiR^{3}AG uses relevant evidence to generate concise, accurate reasoning steps and offer the right answers.

Retrieval-Augmented Generation (RAG) integrates external knowledge into Large Language Models (LLMs) to enhance factuality and reasoning. However, reasoning-augmented LLMs (e.g., OpenAI o1, DeepSeek-R1) introduce substantial computation overhead due to lengthy reasoning traces. To address this trade-off, we conduct a systematic study revealing two reasoning modes in RAG — *Context-Grounded Reasoning* and *Knowledge-Reconciled Reasoning*. Based on these insights, we propose **LiR^{3}AG**, a Lightweight Rerank Reasoning framework that transfers reasoning strategies from large models to smaller non-reasoning models via a three-module design: **Retriever, Reranker, and Reasoning Constructor**. Our approach achieves comparable reasoning accuracy to 32B reasoning models while cutting token and latency costs by over **90%**.

Contributions

- **Systematic Study of Reasoning Strategies.** We conduct the first comprehensive analysis of reasoning behaviors in Retrieval-Augmented Generation (RAG) models, identifying two dominant reasoning modes: *Context-Grounded Reasoning* and *Knowledge-Reconciled Reasoning*.
- **Lightweight Rerank Reasoning Framework (LiR^{3}AG).** We propose a novel framework that transfers reasoning strategies from large reasoning models to smaller non-reasoning models, through three cooperative modules: *Retriever*, *Reranker*, and *Reasoning Constructor*.
- **Efficient and Effective Reasoning Transfer.** LiR^{3}AG enables an 8B non-reasoning model to match or even outperform a 32B reasoning model in multi-hop QA tasks, while reducing generator token usage by up to **98%** and inference latency by **58.6%**.
- **Generalizable and Scalable.** Our framework achieves consistent gains across four multi-hop QA datasets (HotpotQA, 2WikiMultiHopQA, MultiHop-RAG, MuSiQue), demonstrating robust reasoning ability with significantly lower computational overhead.

What Reasoning Models Actually Thinking in RAG

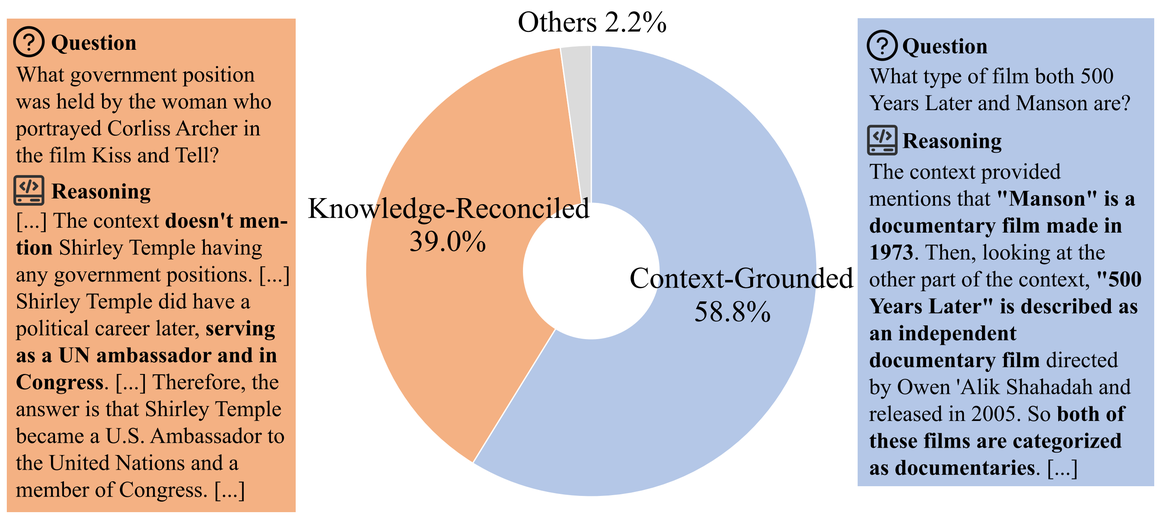

Figure. 2. Distribution of annotated reasoning strategies based on model outputs. The majority of responses follow the Context-Grounded Reasoning strategy (58.8%). Two representative examples are shown to illustrate the feature of each strategy.

- Two dominant reasoning strategies are identified:
  - **Context-Grounded Reasoning**: directly reasoning over retrieved evidence with minimal internal knowledge (58.5%).
  - **Knowledge-Reconciled Reasoning**: verifying or supplementing retrieved information with internal knowledge (39.0%).

Methodology

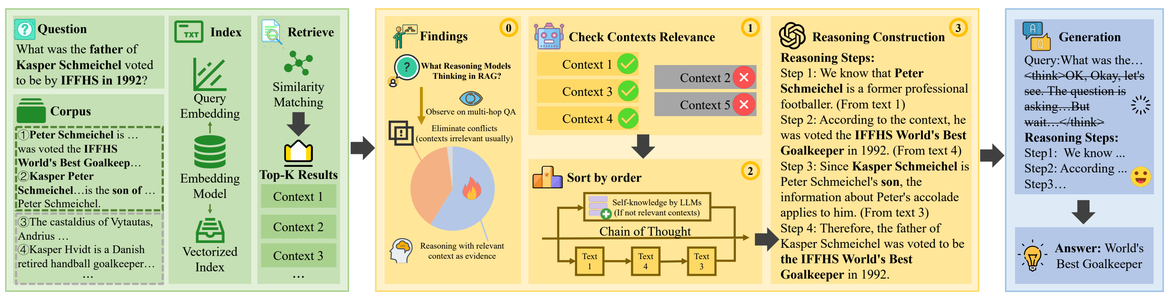

Figure. 3. Overall pipeline of our LiR^{3}AG framework. The Retriever first retrieves potentially relevant contexts. Reranker examines their relevance to the question, filters out irrelevant ones, and orders the remaining contexts according to the expected reasoning sequence. Reasoning Constructor assembles these contexts into structured reasoning steps, which are subsequently passed to the generator to produce the final answer.

Experiments

Method	Backbone LLM	HotpotQA EM	HotpotQA F1	2WikiMultihopQA EM	2WikiMultihopQA F1	MultiHop-RAG EM	MultiHop-RAG F1	MuSiQue EM	MuSiQue F1
Direct Output	8B-no-think	0.172	0.257	0.370	0.382	0.492	0.506	0.038	0.109
	8B-think	0.185	0.285	0.388	0.408	0.529	0.541	0.053	0.143
	14B-no-think	0.150	0.236	0.270	0.276	0.552	0.554	0.024	0.071
	14B-think	0.228	0.316	0.390	0.404	0.546	0.561	0.084	0.174
	32B-no-think	0.176	0.246	0.370	0.398	0.580	0.597	0.032	0.135
	32B-think	0.254	0.358	0.394	0.408	0.608	0.621	0.086	0.190
Vanilla RAG	8B-no-think	0.310	0.403	0.462	0.489	0.602	0.611	0.134	0.239
	8B-think	0.328	0.445	0.562	0.604	0.596	0.613	0.239	0.379
	14B-no-think	0.342	0.441	0.502	0.537	0.628	0.637	0.160	0.277
	14B-think	0.356	0.457	0.562	0.603	0.598	0.615	0.232	0.350
	32B-no-think	0.346	0.452	0.468	0.503	**0.658**	0.668	0.182	0.316
	32B-think	0.380	0.501	0.579	0.634	0.642	0.654	0.271	0.410
Ours	8B-no-think	0.402	0.521	0.586	0.653	0.658	0.673	0.326	0.464

Table 1. The performance of LiR^{3}AG and baseline methods using EM and F1 on four multi-hop QA datasets. **Bold** represents the best results, while underlined denotes the second-best. LiR^{3}AG achieves the SOTA performance even when using the smallest backbone (Qwen3-8B).

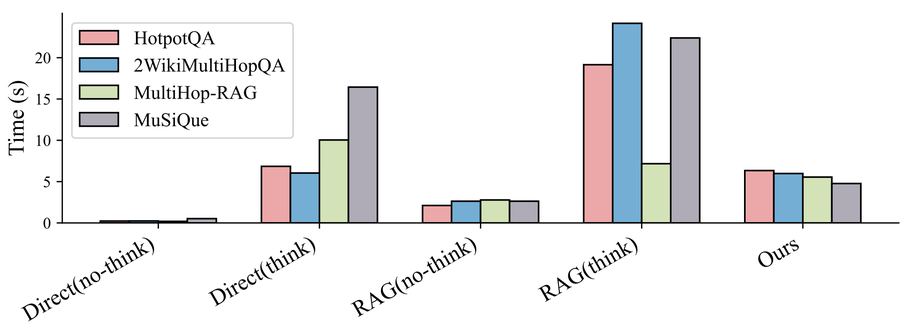

Figure. 4. Time cost of methods on multi-hop QA datasets.

Conclusions: Our approach achieves notable gains in performance while simultaneously reducing the computational costs commonly associated with explicit reasoning, including tokens and inference time.