Faithful in Steps: Improving Generalization and Citation in RAG via Query Decomposition

Yue Liu[1]*, Zhongying Ru[2], Shimin Di[3,4], Jipeng Zhang[1], Ruiyuan Zhang[5], Xiaofang Zhou[1]*

[1]The Hong Kong University of Science and Technology (HKUST)
[2]Independent Researcher
[3]School of Computer Science and Engineering, Southeast University
[4]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications
[5]Hong Kong Generative AI Research and Development Center (HKGAI)
yliumh@connect.ust.hk, zxf@ust.hk

Abstract

Retrieval-augment generation is a prevalent strategy to mitigate hallucinations of LLMs. The attributable RAG (RAGQ) generates quotes for its answers. The quotes indicate which input contexts support the RAG to derive the answers, enhancing the answer's verifiability and trustworthiness. However, existing RAGQs exhibit significant degradation when dealing with questions that require multi-hop reasoning and multi-modal understanding, suffering from over-citation, implicit entity identification failure, and poor generalization. In this paper, we propose a novel RAGQ framework, namely **QDRAG**. QDRAG breaks down the input question into atomic subquestions to identify the implicit entities. Then, the reranker prunes context distractors to eliminate the downstream over-citation. To facilitate query decomposition, we propose two zero-shot approaches: QD-C and QD-R, which guide the QD MLLM to decompose the question based on context knowledge and retrieval rewards, respectively. One interesting finding is that finetuning on the QD task shows better generalizability compared to directly finetuning on the downstream RAGQ task. Experiments on four multi-modal QA benchmarks demonstrate QDRAG's efficacy in grounding answers and generating faithful citations. The framework significantly outperforms all the baselines on both in-domain and out-of-domain tests, even surpassing Gemini-Pro.

Introduction

Retrieval-augmented generation (RAG) (Abootorabi et al. 2025) is a common method to mitigate hallucinations in large-language models. RAG retrieves external data and grounds its answer on the retrieved contexts. Recent studies (Huang et al. 2024; Song et al. 2025) propose RAG with quotes (RAGQ) to improve answer faithfulness. RAGQ outputs include answers and verifiable source references that indicate which retrieved data support the LLM's outputs. Users can verify answers via cited sources, and this alignment between responses and sources is key to answer integrity.

Despite recent advancements, current RAGQs still face challenges for multi-hop and multi-modal questions.

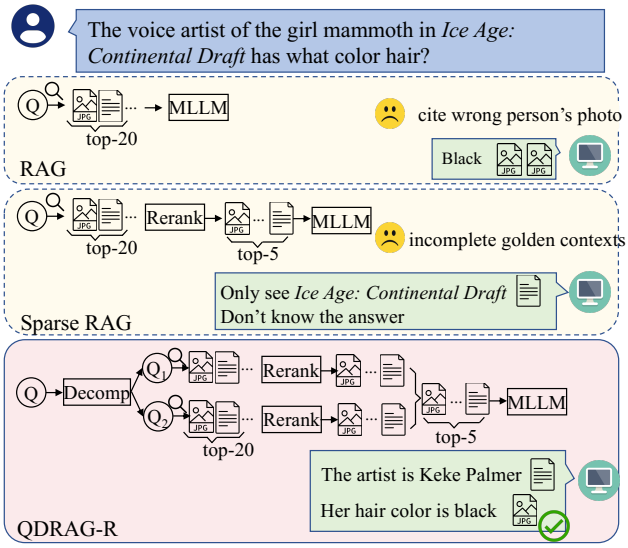


Figure 1: The massive retrieved data in RAG misleads the MLLM to cite the incorrect but similar contexts; Sparse RAG drops key contexts in the reranking step for multi-hop questions; QDRAG filters the distractors while maintaining the key contexts for accurate downstream citations.

***C*-1: Over-citation of incorrect contexts**. The RAGQ cites multi-modal contexts where the local segments align with the answer but neglects to evaluate the broader relevance of these contexts. ***C*-2: Implicit entity retrieval failure**. The post-hoc retrieval approaches (Gao et al. 2023a) rely on retrievers to search for supporting contexts, but simply using the original text of the question as a query often fails due to the key entities being absent in the multi-hop question. ***C*-3: Poor generalization ability.** Instruction fine-tuning methods (Huang et al. 2024; Song et al. 2025) finetune LLMs for downstream tasks. They build training data to enhance LLMs' attribution capability, leading to difficulties in maintaining reference accuracy for dynamic multimedia knowledge (Bi et al. 2024).

One intuitive solution is to use sparse RAG (Zhu et al. 2025b) to pre-filter contexts to a shorter list before MLLM

*Corresponding authors.

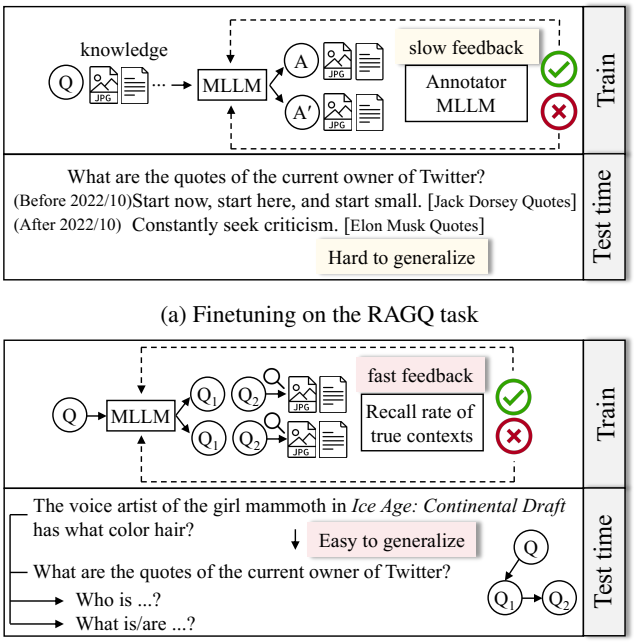
generation, which assesses context relevance in parallel. However, for multi-hop questions, sparse RAG misjudges the contexts with partial knowledge as "irrelevant" due to $\mathcal{C}$**-2**. Another research line decomposes the question into a reasoning plan by few-shot in-context learning. The domain and difficulty gaps between examples and real questions make the MLLM hard to learn how to effectively simplify the multi-hop question (Zhou et al. 2023a; Khot et al. 2023).

To address the above challenges, we propose a novel RAGQ framework named QDRAG, which consists of three main modules: query decomposer (QD), context reranker, and generator. Different from the static contexts used in existing RAGQs, QDRAG filters distractors before answer generation. Sparse contexts help MLLMs focus on correct contexts and generate better quotes ($\mathcal{C}$**-1**). The QD-reranking paradigm assesses context relevance to single-hop subquestions, thereby resolving $\mathcal{C}$**-2**. For query decomposition, we propose zero-shot QD-C and QD-R, two new QD models that eliminate manually written in-context examples. QD-C breaks the QD task into two clearly defined operators: entity replacement and open decomposition. QD-C is plug-and-play for existing RAGs. QD-R uses the recall rate of golden contexts as the reward of a decomposition plan to perform direct preference optimization on the QD task. The retriever feedback significantly boosts the automatic annotation process compared to LLM feedback used by RAGQs. Moreover, the inputs of the decomposition task (a multi-hop question) are context-agnostic and answer-agnostic. The finetuned QDRAG thus does not suffer from the massive and dynamic multimedia contexts. The existence of common decomposition logic among multi-hop questions makes QDRAG more generalizable than existing RAGQs ($\mathcal{C}$-3). Our contributions are summarized as:

- We highlight the close relationship between context sparsity and MLLM context reference capability. It motivates our new QD-Reranking-based RAGQ pipeline.
- We propose two query decomposition methods to identify implicit entities on multi-hop questions, and subquestion-based context reranking to address the over-citation problem on multi-modal contexts. We enhance the model generalizability by finetuning QDRAG on the QD task.
- Extensive experiments demonstrate that QDRAG reduces answer and citation hallucinations and demonstrates remarkable generalization across different base models.

Related Work

RAG with Quotes (RAGQ). The training-free RAGQs use post-hoc retrieval (Gao et al. 2023b), chain-of-thought prompting (Ji et al. 2024) or dynamic programming (Choi et al. 2025) to select supporting documents. For training-based methods, (Asai et al. 2024; Xia et al. 2025) perform supervised-finetuning on RAG LLMs to enable context citations. (Huang et al. 2024; Song et al. 2025) employs reinforcement learning with human feedback or direct preference optimization. Self-Cite (Chuang et al. 2025) utilizes the change of LLM's confidence on an answer when removing the cited documents to calculate the context's importance. In contrast to previous work, QDRAG addresses the RAGQ task



(a) Finetuning on the RAGQ task

(b) Finetuning on the QD task

Figure 2: QDRAG finetunes the MLLM for the QD task, leveraging its fast training and better generalizability.

by performing query decomposition and context filtering to reduce the citation candidates and hallucinations.

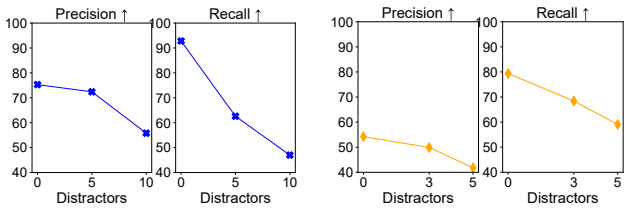
LLM for Reranking. There are four categories in this research line: pointwise, pairwise, listwise, and setwise. The pointwise approach evaluates context relevance one by one. The score can either be a binary label "yes/no", or the normalized likelihood of generating a "yes" response (Zhu et al. 2025b; Yu et al. 2024). Listwise methods prompt an LLM to output the ordered sequence (Sun et al. 2023). Pairwise prompts let the LLM decide which of two contexts is more relevant (Qin et al. 2024). Setwise approach directly outputs the relevant set of contexts (Zhuang et al. 2024; Yu et al. 2024). Among the four categories, the pointwise method is the only one that is parallelizable and also has comparable performance with others (Zhuang et al. 2024).

LLM for Query Decomposition. Least-to-Most (Zhou et al. 2023b), Decomp (Khot et al. 2023), Dynamic Least-to-Most (Drozdov et al. 2023), and Self-Ask (Press et al. 2023) use in-context demonstrations to iteratively decompose the question and answer the subquestions step by step. Recently, SearChain (Xu et al. 2024) adopts the decomposition plan as the generation unit and iteratively refines the plan, surpassing PS (Wang et al. 2023) on multi-hop question answering.

Methodology

Preliminaries

Vanilla Workflow of RAGQ. Given a user query Q, the multi-modal RAGQ retrieves visual and text knowledge pieces from external databases. Then, it merges the knowl-



(a) CIRCO (b) CIRR

Figure 3: Evaluate MLLM's capability to cite images from a multi-modal statement on two datasets: CIRCO and CIRR.

edge pieces with Q as the input context. Next, a multi-modal large language model (MLLM) is prompted to produce an answer with in-line citations in the form of [1][2]. The goal of RAGQ (RAG with quotes) is to ensure that the answer is supported by the cited knowledge sources.

Multi-hop Question Answering. Answering a multi-hop question needs several reasoning steps (Trivedi et al. 2022). The question can be decomposed repeatedly until it can be answered by a single external document. A subquestion Q_i can depend on the answer to another subquestion Q_j. In Figure 1, if Q_i = "Who is the voice artist?", and Q_j = "$\langle$Answer_of_$Q_i\rangle$ has what color hair?", then Q_j relies on Q_i. There are also independent subquestions. For instance, the question "Which film was released first: $\langle a\rangle$ or $\langle b\rangle$?" has two independent subquestions: Q_i = "When was $\langle a\rangle$ released?" and Q_j = "When was $\langle b\rangle$ released?".

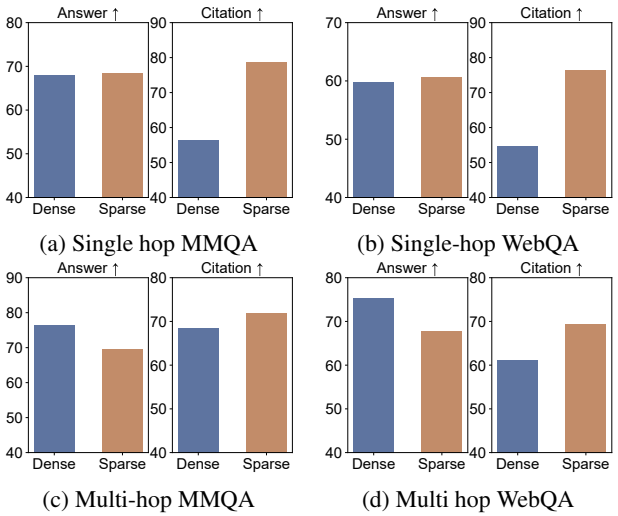
Sparse Context Benefits RAG Quotes

We explore the impact of context sparsity on MLLM's reference performance through two tasks. Our findings show that the sparse RAGs generate better citations than dense RAGs.

The first task is to cite images from an expression of interleaved images and texts. It simulates the scenario that a RAGQ is about to generate citations after the answers. Specifically, the input template is "Compared with the image I_R, the image has $\langle a\rangle$" and a list of images "[1] Image; [2] Image ...". The outputs are image IDs (e.g., [1][3]) that support the statement. We test on two image retrieval datasets (Baldrati et al. 2023; Liu et al. 2021). In Figure 3, the results show that reducing input distractors helps RAGQ reduce the risk of citing wrong images (higher precision) and cite a more complete context set (higher recall). Therefore, sparse RAG is a promising strategy to enhance RAG citations.

The second experiment compares a state-of-the-art sparse RAG (Zhu et al. 2025b) with dense RAG (w/o context filtering) on simultaneously answering the question and citing the multi-modal contexts. Figure 4a and 4b show that the sparse RAG produces better answers and citations than dense RAG on single-hop questions. However, in Figure 4c and 4d, directly applying sparse RAG for multi-hop questions leads to lower answer quality. It implies that sparse RAG drops golden contexts and suffers from hallucinations.

Discussion. We emphasized that sparse RAGs generate better quotes than dense RAGs. We found that the existing sparse RAGs (a.k.a. reranking models) excel in single-hop question answering and citation generation, but face challenges for multi-hop questions. Thus, we propose a novel RAGQ, namely QDRAG, which is superior in both single-hop and multi-hop scenarios.



(a) Single hop MMQA (b) Single-hop WebQA

(c) Multi-hop MMQA (d) Multi hop WebQA

Figure 4: Dense RAG vs Sparse RAG in terms of answer accuracy and attribution capability. **(a)**,**(b)**: Sparse RAG shows superiority on both metrics; **(c)**,**(d)**: Sparse RAG still excels on attribution performance, but has worse answer quality.

QDRAG Overview

As shown in Figure 1, QDRAG comprises three main modules: query decomposer, context reranker, and answer generator. The decomposer derives multiple subquestions, and the reranker purifies the retrieved data by their relevance to subquestions. Afterwards, the answers and citations are generated auto-regressively by the MLLM.

QD-C: Context-Guided Question Decomposition

QD-C defines two operators for query decomposition: entity replacement and open decomposition for inter-dependent and independent subquestions, respectively.

Entity Replacement. We instruct the MLLM to identify any referential expressions (e.g., "The voice artist of the girl mammoth in *Ice Age: Continental Draft*") within the question. Then, the MLLM utilizes the external knowledge to substitute references with entities (e.g., "Keke Palmer"). The operator produces two subquestions for each replacement. One is "Is $\langle$reference$\rangle$ equivalent to $\langle$entity$\rangle$?", the other is the new question after substitution.

Open Decomposition. The MLLM is tasked with allocating queried entities to multiple subquestions to enable answering with a single context. For instance, for question "Which one released first, A or B?", the operator produces two subquestions regarding the released year of two entities. This approach safeguards against the reranker disregarding contexts that contain partial information (e.g., the release year) without explicit comparison.

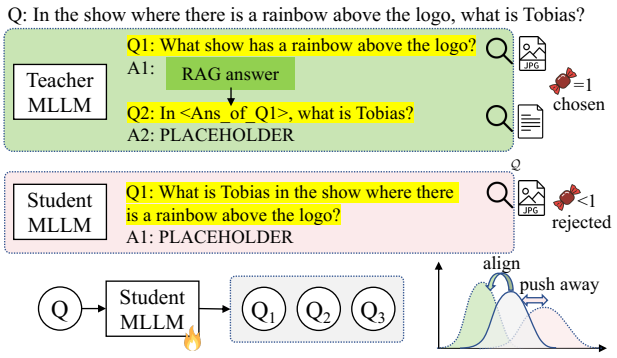


Figure 5: QD-R samples multiple MLLM outputs and uses the recall of golden contexts as reward to train the model.

QD-C eliminates manually-written in-context examples by the well-defined operators. Moreover, the two operators can be implemented into a single MLLM inference for each multi-hop question, which makes QD-C cost-effective and plug-and-play.

QD-R: Efficient and Generalizable QD Alignment

Motivation. Similar to other single-round retrieval RAGs, QD-C assumes that the retrieved documents contain sufficient information to decompose the question. However, a single round of lightweight retrieval often fails to recall the complete set of golden contexts (Yu et al. 2024), which leads to cascading errors that degrade answer and citation quality.

QD-R. We propose retrieval-guided QD (QD-R), a new query decomposition method with multi-round retrievals. Firstly, QD-R takes only the query as input and performs zero-shot query decomposition. The output of QD-R is simply a decomposition plan comprising multiple subquestions. QD-R allows the special token "⟨Ans_of_Qi⟩" in subquestions, which means that the content of the subquestion depends on the answer of previous subquestions. Then, QD-R resolves the placeholders via RAG and replaces the special tokens with the answers to that question. Finally, different from QD-C, each resolved subquestion retrieves its own set of contexts for subsequent reranking.

QDARF. QD-R regularizes its decompositions by **A**ligning MLLM from **R**etrieval **F**eedback (QDARF). As shown in Figure 5, we sample multiple decomposition plans from a teacher MLLM (Qwen2.5VL-32B) and a student MLLM (Qwen2.5VL-7B). The reward of a decomposition plan is defined as the recall of golden contexts by all its subquestions. We adopt the direct preference optimization (DPO) loss function to finetune the MLLM:

$$\mathcal{L} = -E_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \beta\log\frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)}\right)\right], \quad (1)$$

where y_w are y_l are the positive (winning) and the negative (losing) decomposition plans, respectively. The decompositions with the *highest* reward values are deemed as positive, while those with the *lowest* reward values are treated as negative. π_{ref} is the initial MLLM and π_θ is the MLLM to optimize. Optimizing Eq. (1) increases the margin between the positive and negative plans. Consequently, the subquestions would have a larger recall rate of golden contexts after finetuning, which benefits the downstream reranking task.

Context Reranking

QDRAG evaluates the context's relevance to each subquestion and the raw question. The relevance score of a context C to a subquestion Q_i is defined as the probability of "Yes" as the next token under the reranking prompt R:

$$\mathcal{S}(C|Q_i) = \mathcal{P}(\text{"Yes"}|Q_i, C, R)$$

Each subquestion retains a certain number of top-relevant contexts.

Adaptive Budget Allocator

Given a budget N, how to distribute the number of reserved contexts for each subquestion? We propose an entropy-based method. The idea is that higher entropy implies that the MLLM is less certain about selecting the best supporting document, hence requiring more contexts to be retained for that subquestion. Formally,

$$\mathcal{P}(C|Q_i) = \frac{\exp(\mathcal{S}(C|Q_i))}{\sum_{C'\in\mathcal{C}}\exp(\mathcal{S}(C'|Q_i))}$$

$$\mathcal{E}(Q_i) = -\sum_{C'\in\mathcal{C}}\mathcal{P}(C'|Q_i)\log\mathcal{P}(C'|Q_i)$$

Then, the number of reserved contexts allocated to Q_i is

$$n(Q_i) = N\cdot\frac{\mathcal{E}(Q_i)}{\sum_{Q_j\in\mathcal{Q}}\mathcal{E}(Q_j)}$$

Experiment

Experimental Setup

Datasets. We conduct evaluations on four multi-modal question answering datasets: MMQA (Talmor et al. 2021), WebQA (Chang et al. 2021), InfoSeek (Chen et al. 2023) and E-VQA (Mensink et al. 2023). MMQA and WebQA consist of 1,196 and 4,967 testing queries in text form, respectively. Each query necessitates one or more image or text fragments for a complete answer. InfoSeek and E-VQA questions task the models with entity recognition in images and retrieval of relevant textual information for answering.

Evaluation Metrics. We evaluate the RAG outputs from two dimensions: answer quality and context reference faithfulness. For answer quality, we use exact match (EM) and scores graded by LLM-judge. EM measures the ratio of keywords in the ground-truth answer that are exactly mentioned in the model's response. LLM-judge prompts an MLLM ψ to grade the correctness and comprehensiveness of RAG outputs given the golden answers. For context reference capability, following existing work (Song et al. 2025), we perform AutoAIS evaluation using an off-the-shelf MLLM ψ. The inputs to ψ include the user question, the model-generated

Model	Model Size	In-domain						Out-of-domain					
		MMQA			WebQA			InfoSeek			E-VQA		
		EM	LLM	Cite	EM	LLM	Cite	EM	LLM	Cite	EM	LLM	Cite
Retrieval-Generation													
Gemini-Pro	-	43.2	34.2	**68.4**	55.1	26.6	64.6	9.8	0.0	0.9	6.1	3.4	23.6
Gemini-Pro$_{\text{CLIP}}$	-	-	-	60.4	-	-	57.9	-	-	0.4	-	-	27.9
Qwen2.5VL	7B	42.4	40.6	38.2	57.8	41.5	33.0	42.5	36.7	59.6	10.4	7.8	18.5
Qwen2.5VL$_{\text{CLIP}}$	7B	-	-	47.2	-	-	35.5	-	-	58.4	-	-	26.5
InternVL3	14B	49.1	50.1	50.9	58.3	43.6	44.5	36.0	30.5	52.2	8.5	7.3	15.9
InternVL3$_{\text{CLIP}}$	14B	-	-	47.6	-	-	39.7	-	-	52.2	-	-	16.6
Retrieval-Reranking-Generation													
CLIP	428M/7B	35.0	32.4	39.5	50.9	22.9	36.8	40.7	32.4	56.0	9.2	6.5	16.1
MMEmbed	7B	43.6	41.2	51.4	55.9	36.7	44.7	41.8	35.6	60.9	9.5	6.9	22.0
Sparse-RAG	7B	45.9	46.1	48.1	57.6	46.1	39.0	44.5	35.1	58.9	14.6	11.6	15.7
RagVL	13B/7B	42.3	40.8	50.9	56.1	38.9	43.4	-	-	-	-	-	-
Retrieval-QD-Reranking-Generation													
LlamaIndex	7B	43.3	42.1	50.0	56.4	40.8	44.1	10.4	2.1	3.2	13.1	10.7	20.0
QDRAG-C (ours)	7B	43.3	42.2	52.6	56.7	40.2	45.8	42.2	35.4	62.1	9.6	7.5	21.2
QDRAG-C (ours)	14B	46.8	46.9	55.8	57.4	39.7	48.6	36.9	28.7	56.5	6.5	5.5	15.3
QD-Retrieval-Reranking-Generation													
Self-Ask	7B	42.9	41.9	49.9	56.2	40.1	43.8	42.7	34.5	60.6	9.7	8.4	21.1
SearChain	7B	43.7	42.6	49.3	56.3	40.6	45.1	42.4	34.3	64.0	10.0	8.0	21.7
QDRAG-R (ours)	7B	49.2	46.3	57.5	61.7	48.0	59.9	**46.0**	**38.3**	**72.9**	**16.2**	**12.8**	33.8
w/o DPO	7B	46.7	45.4	53.8	61.0	48.3	56.8	42.0	35.2	66.3	13.4	10.6	28.9
QDRAG-R (ours)	14B	**53.1**	**52.8**	65.6	**62.6**	**53.5**	**70.9**	39.8	32.4	68.7	15.7	12.3	**40.4**

Table 1: Overall in-domain and out-of-domain test, where best and second-best results are highlighted in **bold** and underlined, respectively. The retriever for all models is CLIP-ViT-L/14@336px. Single model size means that QD/Reranking/QA use the same MLLM, while multiple model sizes (e.g., 13B/7B) means the reranker model size and the QD/QA model size, respectively.

	MMQA			WebQA		
	EM	LLM	Cite	EM	LLM	Cite
RAG	67.7	64.1	51.3	**70.7**	**64.0**	60.9
MMEmbed	69.7	65.1	71.5	68.0	57.1	69.4
SparseRAG	67.9	66.1	71.9	68.5	60.9	65.3
LlamaIndex	68.0	65.7	69.9	69.6	61.2	72.1
QDRAG-C	**70.7**	**66.4**	**74.7**	**70.7**	63.7	**75.1**

Table 2: Comparison of single-round RAGs with 7B retriever

response, and the cited contexts. The AutoAIS output is a True/False label indicating whether the aggregated contexts support the model response. In this paper, we set ψ as Qwen2.5VL-32B (Bai et al. 2025).

Baselines. As shown in Table 1, the main experiment contains four categories of baselines. The first category comprises dense RAGs with recent MLLM backbones, including the commercial Google Gemini2.5-Pro (Kavukcuoglu 2025) and the open-sourced Qwen2.5-VL (Bai et al. 2025) and InternVL-3 (Zhu et al. 2025a). Following (Huang et al. 2024), this category also contains post-hoc retrieval baselines that select reference contexts with the CLIP retriever. The second category uses multi-modal rerankers to implement sparse RAGs, including CLIP (Radford et al. 2021), RagVL (Chen et al. 2024), MMEmbed (Lin et al. 2025), and SparseRAG (Zhu et al. 2025b). The third category is QD models with single-round retrieval, including LlamaIndex[1] and the proposed QDRAG-C. The last category compares the query decomposition methods under the QDRAG-R pipeline. The baselines include Self-Ask (Press et al. 2023) and SearChain (Xu et al. 2024). In ablation study, we further compare QDRAG with rewriting models (Mao et al. 2024) and the instruction-finetuned RAGQs (Huang et al. 2024).

Implementation Details. If not specified, the backbone of 7B RAGs and 14B RAGs are Qwen2.5-VL-7B (Bai et al. 2025) and InternVL3-14B (Zhu et al. 2025a), respectively. The reranking models reserve top-5 contexts. We adopt low-rank adaptation (LoRA) (Hu et al. 2022) when finetuning the MLLM. The model inference is implemented using the vLLM v0.9.2 framework (Kwon et al. 2023). The experiments are conducted on a Ubuntu server with 2 $\times$ A100 80B NVIDIA GPUs. For detailed hyperparameters and prompts, we refer the reader to the appendix.

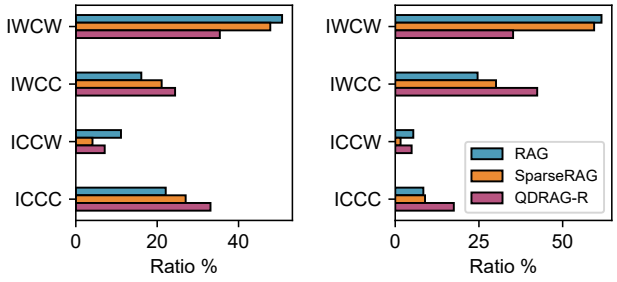
Main Results

Context Filtering Boosts Citation Quality. As shown in Table 1, reranking RAGs consistency show improvements over both prompt-based and post-hoc retrieval baselines.

[1]https://github.com/run-llama/llama_index

	MMQA			WebQA		
	EM	LLM	Cite	EM	LLM	Cite
RaFe-C	43.3	42.2	49.9	55.6	40.3	44.4
QD-C	43.3	42.2	52.6	56.7	40.2	45.8
RaFe-R	43.6	42.0	49.3	55.7	38.7	42.3
RaFe-R_{DPO}	43.8	42.7	50.2	57.4	41.7	48.4
QD-R	46.7	45.4	53.8	61.0	**48.3**	56.8
QD-R_{DPO}	**49.2**	**46.3**	**57.5**	**61.7**	48.0	**59.9**

Table 3: Compare QD with rewriting models



(a) MMQA (b) WebQA

Figure 6: Failure case analysis. **IWCW** ↓: Input wrong context and cite incorrectly; **IWCC** ↑: Input wrong context but cite correctly; **ICCW** ↓: Input correct context but cite incorrectly; **ICCC** ↑: Input correct context and cite correctly.

This highlights our proposition that the sparse context helps MLLMs better attribute their answer to supporting contexts. Comparing CLIP with other rerankers reveals that relying solely on the retrieval score for reranking is suboptimal.

QDRAG-R Achieves Significant Improvements on In-domain and Out-of-Domain Tests. Our QDRAG-R models are trained on the training splits of MMQA and WebQA. InfoSeek and E-VQA have image inputs in all user queries, while MMQA and WebQA only contain pure-text queries. The RAG knowledge bases are also different. Therefore, our training data exhibits out-of-domain characteristics. Nonetheless, QDRAG-R demonstrates a substantial performance advantage on both answer quality and attribution capability over existing reranking RAGs and QD-reranking RAGs. The QDRAG-R-14B outperforms the proprietary MLLM Gemini2.5-pro in most cases.

QDRAG-C Excels with More Powerful Retrievers. In Table 1, QDRAG-C shines at resource attribution but is less splashy for answer quality. This highlights our proposition that QDRAG-C is potentially sensitive to initial retrieval quality. In Table 2, we replace the lightweight CLIP retriever with MMEmbed-7B. Then QDRAG-C demonstrates improvements in both answer quality and attribution capability.

Analysis

Rewriting-R vs QD-R Rewriting is another important query processing strategy. We compare QD-reranking with the recent query rewriting method RaFe (Mao et al. 2024) by adapting it for RaFe-reranking. As shown in Table 3, in single-round retrieval setup, QD-C shows improvement compared to RaFe-C. In multi-round retrieval setup, we construct the RaFe training dataset similar to that of QD-R. The only difference is that RaFe instructs the MLLM to rewrite the query. The results show that query decomposition is indeed an essential step for RAG trustworthiness and answer quality.

Pipe	FT task	InfoSeek		EVQA	
		LLM	Cite	LLM	Cite
RAG	-	36.7	59.6	7.8	18.5
	RAGQ	32.9	55.6	6.4	18.7
	QD-R	36.6	62.3	7.8	19.1
QDRAG-R	-	35.2	66.3	10.6	28.9
	RAGQ	33.0	68.4	8.6	26.0
	QD-R	**38.3**	**72.9**	**12.8**	**33.8**

Table 4: Out-of-domain test: Front vs QDRAG

FT task	Feedback	Auto	Cost
RAGQ	LLM-judge	✓	28h
QD-R	Retrieval	✓	2h

Table 5: Cost of feedback when building training data

Finetuning Task We compare QDRAG with a representative RAGQ finetuning method: Front (Huang et al. 2024). Front retrieves contexts for training questions and uses a teacher model and a student model to generate positive and negative answers with citations, respectively. For fair comparison, we use the same teacher model and student model as QDRAG. The results are shown in Table 4. Similar to published results (Huang et al. 2024; Song et al. 2025), we observe descending answer accuracy of RAGQ on out-of-domain datasets. QDRAG shows a significantly better generalization capability on out-of-domain datasets.

In Table 5, we compare the annotation cost of Front and QDRAG. The results show that the retrieval feedback is significantly more efficient than MLLM feedback.

What Makes QD-R Better? In Figure 6, we divide the inputs into four cases for each model. Our key findings are: (1) most incorrect references are caused by incorrect context inputs, which highlights the importance of QD-based context reranking; (2) QD-R significantly reduces the failure cases.

Ablation Study We conduct a set of ablations of QDRAG to identify which factors play key roles. We examine the benefits of the replacement (R) and the decomposition (D) operators in QD-C. From Table 6, we observe that both of them are critical for effective query decomposition. Compared with no query decomposition, QD-C shows a large performance gain. We test the contributions of QD-R training datasets in Table 7. The result shows QD-R's scalability w.r.t. the training data size. In both models, the adaptive context number assignment module (S) provides performance gains.

Efficiency & Cost In Figure 8a, we analyze the efficiency and scalability w.r.t. the number of retrieved contexts at each

D	R	S	MMQA EM	MMQA LLM	MMQA Cite	WebQA EM	WebQA LLM	WebQA Cite
✓	✓	✓	43.3	42.2	52.6	56.7	40.2	45.8
✓		✓	43.3	41.8	51.0	56.8	40.8	45.4
	✓	✓	43.8	41.4	50.9	56.9	39.4	45.2
✓	✓		44.0	42.0	52.0	56.5	40.7	45.8
			43.5	42.0	48.8	55.8	39.8	41.9

Table 6: Ablation on QDRAG-C

M	W	S	MMQA EM	MMQA LLM	MMQA Cite	WebQA EM	WebQA LLM	WebQA Cite
✓	✓	✓	49.2	46.3	57.5	61.7	48.0	59.9
✓		✓	46.9	44.6	56.9	61.1	48.1	56.9
	✓	✓	46.4	45.4	54.3	61.6	48.3	58.3
✓	✓		48.8	46.0	56.5	61.9	47.8	60.6
			46.9	45.0	54.8	61.0	48.1	56.9

Table 7: Ablation on QDRAG-R Traing Data

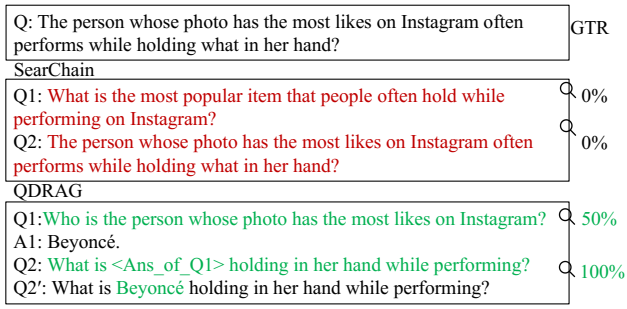
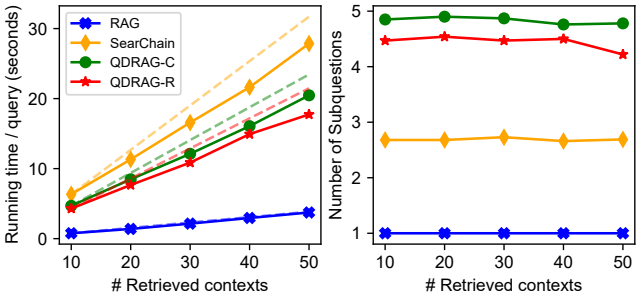


Figure 7: Case study: QDRAG generates better query decompositions. GTR: Golden context Recall.

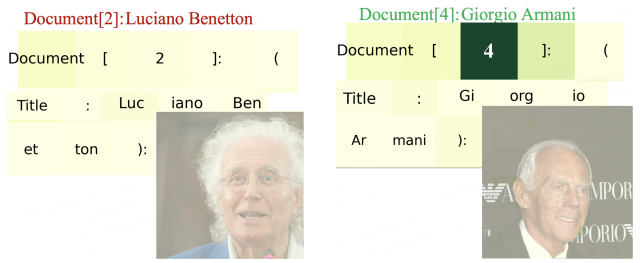


(a) Analysis of efficiency and scalability

Method	$\#n$	$\#m$	$\#r$	Cost ↓
SearChain	25715	391	10	0.070
QDRAG-C	4810	451	1	0.019
QDRAG-R	12660	237	4	0.035

(b) Analysis of QD costs per query in dollars

Figure 8: Efficient & cost analysis



(a) Dense context RAG (b) QDRAG

Figure 9: Visualization of image contexts with the largest attention values when MLLM generates the references after "What color is Giorgio Armani's hair? White. According to Document ␣"

retrieval round. The dashed lines indicate the linear growth of running time. Compared with baselines, our method substantially boosts task performance without significantly increasing time consumption compared to the baselines. Notably, QDRAGs generate more effective subquestions with less running time.

In Figure 8b, we analyze the costs of query decomposition in dollars. The reported metrics include the average number of input tokens (n), output tokens (m), and MLLM calls (r). The overall cost is calculated by $\sum_r (n_r p_1 + m_r p_2)$, where p_1 and p_2 denote the unit price of input and output tokens of Gemini2.5-pro, respectively. QDRAG-C only incurs a single MLLM call to make it cost-effective. QDRAG-R has a slightly higher cost but has higher performance.

Case Study In Figure 7, the baseline decompositions have almost the same level of difficulty with the input question. Our QDRAG successfully simplifies the question step by step. The baselines prompt an MLLM to decompose the query in the open exploration space. This may lead to unsolvable subqueries or rewritings rather than really reducing the question complexity. In our method, we enable MLLM-IR interaction by preference optimization. The MLLM in QDRAG tends to generate reasonable subquestions that have large rewards, i.e., the recall rate of golden contexts. This benefits downstream tasks of question answering and contextual referencing.

In Figure 9, we visualize the image context attention that has the largest attention values on the context ID (e.g., Document [4]) when the QA MLLM generates references. The dense RAG spends large attention on a photo of an incorrect person who also has white hair, while QDRAG is more confident and wise when generating reference IDs.

Conclusion

In this work, we propose QDRAG, a novel RAG framework to enhance attributed RAG. We present that the sparse context helps LLM concentrate on important contexts and point out the limitation of existing sparse RAGs. We propose two new query-decomposition methods. In particular, QD-C is guided by initial contexts, while QD-R is guided by retrieval rewards. QDRAG successfully addresses the three challenges of existing RAGQs. Extensive experiments are conducted to validate the generalizability and scalability of QDRAG.

Acknowledgments

The research work described in this paper was supported by Hong Kong Research Grants Council (grant# T43-513/23-N, T22-607/24N). It was partially conducted in JC STEM Lab of Data Science Foundations funded by The Hong Kong Jockey Club Charities Trust. Shimin Di's work is supported by National Science Foundation of China (NSFC) under Grant No. 62506075.

References

Abootorabi, M. M.; Zobeiri, A.; Dehghani, M.; Mohammadkhani, M.; Mohammadi, B.; Ghahroodi, O.; Baghshah, M. S.; and Asgari, E. 2025. Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. *arXiv preprint arXiv:2502.08826*.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-VL Technical Teport. *arXiv preprint arXiv:2502.13923*.

Baldrati, A.; Agnolucci, L.; Bertini, M.; and Del Bimbo, A. 2023. Zero-shot Composed Image Retrieval with Textual Inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15338–15347.

Bi, B.; Huang, S.; Wang, Y.; Yang, T.; Zhang, Z.; Huang, H.; Mei, L.; Fang, J.; Li, Z.; Wei, F.; et al. 2024. Context-DPO: Aligning Language Models for Context-Faithfulness. *arXiv preprint arXiv:2412.15280*.

Chang, Y.; Narang, M. B.; Suzuki, H.; Cao, G.; Gao, J.; and Bisk, Y. 2021. WebQA: Multihop and Multimodal QA. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16474–16483.

Chen, Y.; Hu, H.; Luan, Y.; Sun, H.; Changpinyo, S.; Ritter, A.; and Chang, M.-W. 2023. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14948–14968. Singapore: Association for Computational Linguistics.

Chen, Z.; Xu, C.; Qi, Y.; and Guo, J. 2024. MLLM Is a Strong Reranker: Advancing Multimodal Retrieval-augmented Generation via Knowledge-enhanced Reranking and Noise-injected Training. arXiv:2407.21439.

Choi, N.; Byun, G.; Chung, A.; Paek, E. S.; Lee, S.; and Choi, J. D. 2025. Trustworthy Answers, Messier Data: Bridging the Gap in Low-Resource Retrieval-Augmented Generation for Domain Expert Systems. *arXiv preprint arXiv:2502.19596*.

Chuang, Y.-S.; Cohen-Wang, B.; Shen, Z.; Wu, Z.; Xu, H.; Lin, X. V.; Glass, J. R.; Li, S.-W.; and tau Yih, W. 2025. SelfCite: Self-Supervised Alignment for Context Attribution in Large Language Models. In *Forty-second International Conference on Machine Learning*.

Drozdov, A.; Schärli, N.; Akyürek, E.; Scales, N.; Song, X.; Chen, X.; Bousquet, O.; and Zhou, D. 2023. Compositional Semantic Parsing with Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A. T.; Fan, Y.; Zhao, V.; Lao, N.; Lee, H.; Juan, D.-C.; and Guu, K. 2023a. RARR: Researching and Revising What Language Models Say, Using Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16477–16508. Toronto, Canada: Association for Computational Linguistics.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023b. Enabling Large Language Models to Generate Text with Citations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488. Singapore: Association for Computational Linguistics.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, L.; Feng, X.; Ma, W.; Gu, Y.; Zhong, W.; Feng, X.; Yu, W.; Peng, W.; Tang, D.; Tu, D.; and Qin, B. 2024. Learning Fine-Grained Grounded Citations for Attributed Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 14095–14113. Bangkok, Thailand: Association for Computational Linguistics.

Ji, B.; Liu, H.; Du, M.; and Ng, S.-K. 2024. Chain-of-Thought Improves Text Generation with Citations in Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18345–18353.

Kavukcuoglu, K. 2025. Gemini 2.5: Our Most Intelligent AI Model. Accessed: 2025-07-21.

Khot, T.; Trivedi, H.; Finlayson, M.; Fu, Y.; Richardson, K.; Clark, P.; and Sabharwal, A. 2023. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. In *The Eleventh International Conference on Learning Representations*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lin, S.-C.; Lee, C.; Shoeybi, M.; Lin, J.; Catanzaro, B.; and Ping, W. 2025. MM-Embed: Universal Multimodal Retrieval With Multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2125–2134.

Mao, S.; Jiang, Y.; Chen, B.; Li, X.; Wang, P.; Wang, X.; Xie, P.; Huang, F.; Chen, H.; and Zhang, N. 2024. RaFe: Ranking Feedback Improves Query Rewriting for RAG. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP*

2024, 884–901. Miami, Florida, USA: Association for Computational Linguistics.

Mensink, T.; Uijlings, J.; Castrejon, L.; Goel, A.; Cadar, F.; Zhou, H.; Sha, F.; Araujo, A.; and Ferrari, V. 2023. Encyclopedic VQA: Visual Questions about Detailed Properties of Fine-grained Categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3113–3124.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.; and Lewis, M. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711. Singapore: Association for Computational Linguistics.

Qin, Z.; Jagerman, R.; Hui, K.; Zhuang, H.; Wu, J.; Yan, L.; Shen, J.; Liu, T.; Liu, J.; Metzler, D.; Wang, X.; and Bendersky, M. 2024. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 1504–1518. Mexico City, Mexico: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Song, M.; Sim, S. H.; Bhardwaj, R.; Chieu, H. L.; Majumder, N.; and Poria, S. 2025. Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse. In *The Thirteenth International Conference on Learning Representations*.

Sun, W.; Yan, L.; Ma, X.; Wang, S.; Ren, P.; Chen, Z.; Yin, D.; and Ren, Z. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14918–14937. Singapore: Association for Computational Linguistics.

Talmor, A.; Yoran, O.; Catav, A.; Lahav, D.; Wang, Y.; Asai, A.; Ilharco, G.; Hajishirzi, H.; and Berant, J. 2021. MultiModalQA: Complex Question Answering over Text, Tables and Images. In *International Conference on Learning Representations*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10: 539–554.

Wang, L.; Xu, W.; Lan, Y.; Hu, Z.; Lan, Y.; Lee, R. K.; and Lim, E. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 2609–2634. Association for Computational Linguistics.

Xia, S.; Wang, X.; Liang, J.; Zhang, Y.; Zhou, W.; Deng, J.; Yu, F.; and Xiao, Y. 2025. Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 969–988. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.

Xu, S.; Pang, L.; Shen, H.; Cheng, X.; and Chua, T.-S. 2024. Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-Intensive Tasks. In *Proceedings of the ACM Web Conference 2024*, 1362–1373.

Yu, Y.; Ping, W.; Liu, Z.; Wang, B.; You, J.; Zhang, C.; Shoeybi, M.; and Catanzaro, B. 2024. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. *Advances in Neural Information Processing Systems*, 37: 121156–121184.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023a. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023b. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025a. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.

Zhu, Y.; Gu, J.-C.; Sikora, C.; Ko, H.; Liu, Y.; Lin, C.-C.; Shu, L.; Luo, L.; Meng, L.; Liu, B.; and Chen, J. 2025b. Accelerating Inference of Retrieval-Augmented Generation via Sparse Context Selection. In *The Thirteenth International Conference on Learning Representations*.

Zhuang, S.; Zhuang, H.; Koopman, B.; and Zuccon, G. 2024. A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, 38–47. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704314.