RaCoT: Plug-and-Play Contrastive Example Generation Mechanism for Enhanced LLM Reasoning Reliability

Kaitong Cai[1*], **Jusheng Zhang**[1*], **Yijia Fan**[1], **Jing Yang**[1], **Keze Wang**[1,2†]

[1]Sun Yat-sen University
[2]Guangdong Key Laboratory of Big Data Analysis and Processing

Abstract

Retrieval-Augmented Generation (RAG) faces a core bottleneck with knowledge-sparse and semantically ambiguous long-tail queries, where retrieval noise distorts reasoning and necessitates costly post-processing. To tackle this, we propose RaCoT (Retrieval-aware Contrastive-of-Thought), a novel framework that shifts contrastive thinking to the pre-retrieval stage. By automatically generating a semantically adjacent yet differently answered contrastive question and extracting a Δ-Prompt to capture their key differences, RaCoT guides the model to proactively focus on the "critical details that determine answer divergence." This approach allows it to suppress semantic interference within a single retrieval pass, overcoming the theoretical bottleneck of single-vector queries that struggle to simultaneously encode signals for what to attend to and what to ignore. On six authoritative benchmarks, including PopQA and TriviaQA-unfiltered, RaCoT outperforms strong baselines like RankRAG and Self-RAG by 0.9-2.4 percentage points. It exhibits superior robustness, with a performance drop of only 8.6% in adversarial tests, far surpassing the over 15% degradation in other methods. Furthermore, its low latency (3.12s) and token overhead (11.54) place it on the accuracy-efficiency Pareto frontier, while ablation studies validate the necessity of each component. Ultimately, RaCoT reframes the RAG paradigm from "post-hoc context cleaning" to "a priori shaping of discriminative reasoning", offering an efficient and robust path toward reliable AI systems for real-time, resource-constrained deployments.

Introduction

Recently, large language models (LLMs) (Vaswani et al. 2023; Brown and et al. 2020; Bommasani and et al. 2022; Radford et al. 2018; Wang et al. 2024a,b; Li et al. 2025c,b), such as GPT-4 (OpenAI 2024) and the LLaMA (AI@Meta 2024) family, have demonstrated remarkable progress across a wide range of natural language processing tasks, exhibiting strong general-purpose capabilities. However, these models are inherently constrained by their *knowledge cutoff (Bommasani and et al. 2022; Zhao and et al. 2025)*, which renders them less effective when faced with knowledge-sparse and semantically ambiguous long-tail questions. To mitigate this limitation, the Retrieval-Augmented Generation (RAG) (Asai et al. 2024; Chan et al. 2024; Kim et al. 2024; Lewis and et al. 2021; Guu et al. 2020; Liu et al. 2024, 2025; Zhang et al. 2025a,c) paradigm enhances factual accuracy by incorporating external knowledge sources. A straightforward way to improve performance in this framework is to increase the retrieval scope, i.e., (Izacard and Grave 2021), retrieving more documents in hopes of covering potential answers through redundancy.

While this approach has shown performance gains in certain scenarios, it faces significant challenges in the context of long-tail reasoning (Fu et al. 2023; Zhang et al. 2019; Sun et al. 2023; Zhang et al. 2023; Liu et al. 2019; Zhang et al. 2025h). Specifically, when dealing with vague or ambiguous queries, expanding the retrieval set often introduces numerous surface-level but semantically irrelevant documents, which dilute the model's attention and may even lead to catastrophic degradation in answer quality. Prior work has shown (Huang et al. 2021; Zhang et al. 2023; de Alvis and Seneviratne 2024; Zhang et al. 2025b) that excessive noise in retrieved documents can lead to significant drops in model performance once it surpasses a certain threshold. To compensate for such noise, recent methods such as Self-RAG (Asai et al. 2024)and RankRAG (Yu et al. 2024) resort to complex post-retrieval strategies, like re-ranking or reflective filtering, to extract more relevant context. However, these mechanisms substantially increase the computational cost and inference latency, thereby limiting their practicality in real-world deployments.

These challenges point to a more fundamental limitation: existing RAG systems predominantly operate via *passive filtering* and lack an intrinsic ability to discriminate among retrieved information during reasoning (Gupta, Ranjan, and Singh 2024; Liu et al. 2023a). This leads us to the following core research question:

> (Q): *Can we design a mechanism that does not merely optimize input content post hoc, but fundamentally enhances the model's discriminative attention over retrieved information, enabling it to actively focus on semantically critical evidence in the presence of noise and ambiguity?*

To address this question, we shift the focus from "optimiz-

*These authors contributed equally.
†Corresponding author: kezewang@gmail.com

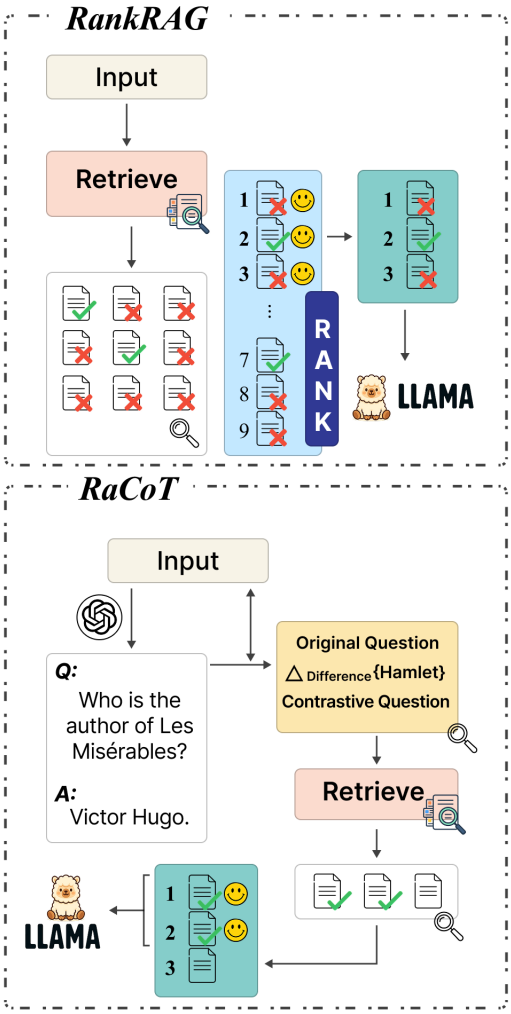

Figure 1: RankRAG improves retrieval quality by ranking results based on the original question, while our RaCoT generates contrastive questions to guide more reasoning-oriented retrieval.

ing context quality" to "systematically guiding the model's reasoning process". Rather than repeatedly filtering out irrelevant documents to construct a noise-free context, we propose to train the model to *reason robustly in realistic (Lewis et al. 2021; Liu et al. 2023b; Zhang et al. 2025i; Tang et al. 2025), noisy retrieval settings*, by actively identifying and grounding on key evidence during a single-pass retrieval.

Inspired by how human experts often resolve complex problems through comparative analysis (Tversky 1977; Eiter et al. 2023; Yan et al. 2024; Zhang et al. 2025g), we propose a novel inference paradigm rooted in *contrastive thinking*. We first validate an intuitive hypothesis: providing a high-quality positive example during retrieval can guide the model toward better document selection. Building on this, we propose the **RaCoT** (Retrieval-aware Contrastive-of-Thought) framework. Its core idea is to go beyond mere positive prompting by introducing explicit *contrastive signals* through the construction of a semantically adjacent yet distinct *contrastive question* and a corresponding *difference prompt* (Δ). By jointly presenting the target and contrastive questions alongside their critical semantic differences, the model is encouraged to focus on *which details lead to fundamentally different answers*, thereby inducing a more discriminative and robust attention mechanism. The **main contributions** of this work can be summarized as follows: i) We present a systematic analysis of the limitations in existing RAG methods when applied to long-tail and ambiguous queries. In particular, we identify key inefficiencies and effectiveness bottlenecks caused by reliance on post-retrieval filtering or the use of single positive exemplars; ii) We introduce a novel framework, **RaCoT**, which systematically applies *contrastive reasoning* to the pre-retrieval stage to enhance query representation. By constructing contrastive exemplars prior to retrieval, RaCoT enhances the semantic discriminability of retrieval intents in a single-pass setup; iii) Extensive experiments across multiple benchmark datasets demonstrate that RaCoT significantly improves both retrieval quality and answer accuracy, especially in scenarios involving noisy and ambiguous queries. The framework consistently outperforms existing methods in terms of both effectiveness and robustness.

Related Works

The RAG Paradigm and Its Inherent Limitations. Recently, large language models (LLMs) like GPT-4 (OpenAI 2024) and the LLaMA (AI@Meta 2024; Grattafiori 2024; Zhang et al. 2025d)series have shown remarkable general capabilities. However, their knowledge is frozen at a specific point in time, leading to factual errors or "hallucinations" when handling knowledge-sparse or semantically ambiguous long-tail queries (Bubeck et al. 2023; Huang et al. 2021; Liu et al. 2019; Zhang et al. 2023, 2025e). The Retrieval-Augmented Generation (RAG) (Chan et al. 2024; Yu et al. 2024; Kim et al. 2024; Gao et al. 2024; Lewis and et al. 2021) paradigm was introduced to mitigate this by grounding LLMs in external knowledge sources. The foundational RAG framework proposed by Lewis et al. (Lewis and et al. 2021) integrates a retriever with a generator, modeling the generation probability as $p(A|Q, d)$, and performs well on simple tasks. Nevertheless, for ambiguous queries, expanding the retrieval scope often introduces irrelevant documents, causing attention dilution and a "catastrophic decline" in performance once noise surpasses a certain threshold.

Post-Retrieval Optimization: A Reactive Approach. To combat retrieval noise, a dominant line of work has focused on **post-retrieval optimization**. These methods apply complex processing after retrieving an initial set of documents. For instance, Self-RAG (Asai et al., 2023) (Asai et al. 2024) introduces a self-reflection mechanism to filter documents, RankRAG (Zhang et al., 2023) (Yu et al. 2024) uses reranking to improve context quality, and methods like IterDRAG (Yue et al. 2025) employ iterative retrieval for multi-hop reasoning. However, these strategies are reactive by nature (Izacard and Grave 2021; Guu et al. 2020). They treat the initial retrieval as a noisy channel that must be cleaned,

Method	Efficiency	Noise Resistance	Multi-Hop
Basic RAG	✓	✗	✗
Self-RAG	▲	✓	▲
RankRAG	▲	✓	▲
IterDRAG	✗	✓	✓
AutoRAG	▲	✗	✗
RaCoT (Ours)	✓	✓	✓

Table 1: Qualitative comparison of related works (✓: strong, ▲: medium, ✗: weak). The evaluation criteria are justified in the Experiments section.

imposing substantial computational and latency overheads. Fundamentally, they are workarounds that do not address the root issue, i.e., a single query vector representation struggles to encode both what to attend to and what to ignore.

From Query Rewriting to Pre-Retrieval Contrastive Enhancement. Recognizing the inefficiency of post-retrieval processing, a more recent trend has shifted towards **pre-retrieval query enhancement**. This line of work, exemplified by methods like AutoRAG(Kim et al. 2024), focuses on *query rewriting* to refine user intent. While these methods mark an important conceptual shift towards proactive enhancement(Chuang et al. 2023; Ma et al. 2023; Xiong and et al. 2021; Fang et al. 2025; Li et al. 2025a; Zhang et al. 2025f), they aim primarily to improve the **"positive" representation** of the query before it is sent to the retriever.

However, they are constrained by a critical limitation: they lack an explicit **"negative" or "contrastive" signal**. They can refine a query for clarity (Zhou and Gong 2022; Qu and Ding 2021), but they cannot proactively arm the model to distinguish correct information from cleverly disguised distractors. This leaves a crucial gap, as even a well-written query can be semantically ambiguous and retrieve misleading documents.

Our work, RaCoT, is designed to fill this specific gap. To our knowledge, **RaCoT is the first framework to systematically introduce an explicit contrastive reasoning mechanism into the pre-retrieval stage**. Instead of merely rewriting the query, RaCoT constructs a **contrastive triplet**, which includes the original question (Q_{target}), a semantically adjacent but differently answered question ($Q_{contrast}$), and a **differential prompt (Δ)** that precisely isolates their key differences. This approach moves beyond simple query optimization by injecting a discriminative signal that teaches the model *what to ignore*, fundamentally enhancing the query's robustness before retrieval occurs.

Qualitative Comparison. Table 1 compares RaCoT with representative RAG methods across key dimensions. It highlights how existing methods often sacrifice efficiency for robustness, whereas RaCoT strikes a superior balance.

- **Efficiency** is benchmarked by latency and token overhead (see Figure 5), where RaCoT is positioned favorably on the accuracy-efficiency Pareto frontier.
- **Noise Resistance** is assessed by the performance drop under adversarial distractor injection (see Figures 3 and 4), where RaCoT shows a minimal 8.6% degradation.
- **Multi-Hop Reasoning** capability is measured on complex QA benchmarks like HotpotQA (see Table 2 and Table 3), where RaCoT demonstrates strong performance.

Methodology

Reformulating the Problem: The Semantic Bottleneck in Query Representation

From a probabilistic perspective, the retrieval-augmented generation (RAG) framework models the posterior probability of the answer $p(A|Q)$ by marginalizing over the retrieved documents d:

$$p(A|Q) = \sum_{d \in \mathcal{C}} p(A|Q, d) \cdot p(d|Q)$$

Here, the generator $\mathcal{M}_{\text{gen}}$ models $p(A|Q, d)$, while the retriever $\mathcal{R}$ estimates $p(d|Q)$. Although this formulation is theoretically sound, the practical performance of RAG systems is often bottlenecked by the retrieval stage.

We argue that the root cause lies in the **inherent insufficiency of query representation**. Whether sparse retrievers (e.g., BM25) or dense retrievers are used, the core operation is to project the original query Q into a fixed representation, such as a bag-of-words vector or a dense embedding $v_Q \in \mathbb{R}^d$. While such representations suffice for simple factual questions, they exhibit severe limitations in long-tail queries involving nuanced semantic ambiguities. A single static vector v_Q struggles to simultaneously encode both "what to attend to" and "what to ignore", two opposing but essential signals for discriminative reasoning. As a result, the similarity computation (e.g., $\text{sim}(v_Q, v_d)$) becomes unreliable, making it difficult for the retriever to distinguish genuinely relevant documents from *semantic distractors* that are topically related but address subtly different questions. This ultimately degrades the quality of the retrieval distribution $p(d|Q)$.

This analysis suggests that retrieval-stage re-ranking serves as a remedy rather than a fundamental solution. It motivates a new research paradigm: *Instead of purifying retrieved documents after the fact, can we enhance the query representation before retrieval?*

RaCoT: Enhanced Discriminative Representations via Contrastive Reasoning

To address the aforementioned representational bottleneck, we propose the RaCoT framework. Rather than directly modifying the retriever $\mathcal{R}$ or the generator $\mathcal{M}_{\text{gen}}$, RaCoT introduces a dynamic, context-aware representation enhancement module, denoted as Π_{RaCoT}. This module transforms the original query Q, which may suffer from representational limitations, into a *discriminatively-enhanced retrieval representation* Q^*:

$$Q^* = \Pi_{\text{RaCoT}}(Q_{\text{target}}, Q_{\text{contrast}}, \Delta)$$

Unlike the original static vector representation v_Q, the enhanced query Q^* is generated through an explicit process of differential reasoning. **In practice, Q^* is a semantically enriched textual object, for instance, a hypothetical sketch of an ideal supporting document,** that encodes

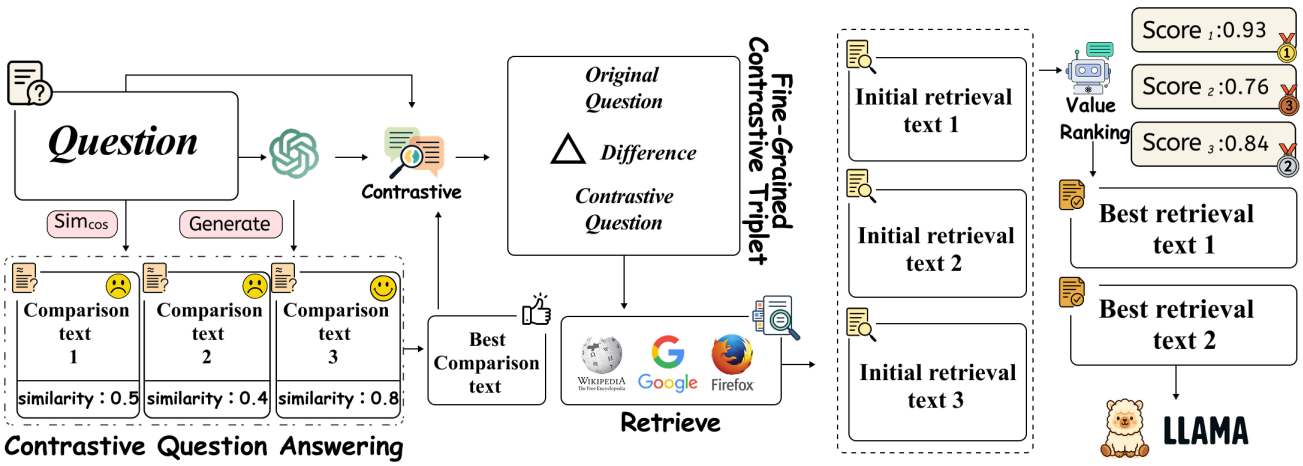

Figure 2: RaCoT enhances complex question answering by generating contrastive questions that differ in key semantics, retrieving evidence for both the original and contrastive questions, and forming fine-grained triplets. A value ranking network then scores and ranks candidate passages to select the most informative ones, improving retrieval-augmented generation with stronger semantic discrimination.

Algorithm 1: RaCoT Inference Pipeline

Require: Target question Q_{target}, corpus $\mathcal{C}$, LLMs $(\mathcal{M}_{\text{teacher}}, \mathcal{M}_{\text{RaCoT}}, \mathcal{M}_{\text{gen}})$, retriever $\mathcal{R}$, parameters K, τ
Ensure: Answer A

Stage 1: Contrastive Sample Generation (Offline)
$Q_{\text{contrast}}, \Delta \leftarrow \mathcal{M}_{\text{teacher}}(\mathcal{T}_{\text{gen}}(Q_{\text{target}}))$

Stage 2: Intent Refinement and Retrieval
$Q^* \leftarrow \mathcal{M}_{\text{RaCoT}}(\mathcal{T}_{\Delta}(Q_{\text{target}}, Q_{\text{contrast}}, \Delta))$
$D_{\text{cand}} \leftarrow \mathcal{R}(Q^*, \mathcal{C}, K)$

Stage 3: One-Pass Filtering and Context Refinement
$\mathcal{C}_{\text{final}} \leftarrow \emptyset$
for all $d_i \in D_{\text{cand}}$ **do**
  $s_i \leftarrow \text{Score}_{\mathcal{M}_{\text{RaCoT}}}(\mathcal{T}_{\text{filter}}(d_i, Q_{\text{target}}, \Delta))$
  if $s_i > \tau$ **then**
    $\mathcal{C}_{\text{final}} \leftarrow \mathcal{C}_{\text{final}} \cup \{d_i\}$
  end if
end for

Stage 4: Contrast-Aware Answer Generation
$A \leftarrow \mathcal{M}_{\text{gen}}(\mathcal{T}_{\text{ans}}(Q_{\text{target}}, \mathcal{C}_{\text{final}}, \Delta))$

return A

meta-information regarding how to handle ambiguity. We hypothesize that the resulting retrieval distribution $p(d|Q^*)$ is significantly more accurate than the original $p(d|Q)$. The complete procedural pipeline is illustrated in **Algorithm 1**.

Core Mechanism: Δ-Prompting

Δ-Prompting serves as the core mechanism for realizing the representation enhancement module Π_{RaCoT}. It enforces contrastive reasoning within the language model to inject discriminative signals into the final retrieval intent.

Construction of Contrastive Samples This step provides the necessary "negative" examples or semantic references to facilitate subsequent differential reasoning. Concretely, for each target question Q_{target}, we utilize a teacher model $\mathcal{M}_{\text{teacher}}$ to generate a high-quality contrastive question Q_{contrast} along with its associated difference description Δ:

$$(Q_{\text{contrast}}, \Delta) = \mathcal{M}_{\text{teacher}}(\mathcal{T}_{\text{gen}}(Q_{\text{target}})) \tag{1}$$

To ensure the quality and semantic relevance of the contrastive samples, which is crucial for the framework's effectiveness, we leverage a powerful, instruction-tuned LLM as the teacher model. Acknowledging this dependency, we further analyze its potential impact in our ablation studies. To ensure the effectiveness of contrast, we employ a constrained optimization procedure to select Q_{contrast}:

$$Q_{\text{contrast}} = \arg\max_{Q'_i} \ \text{sim}_{\cos}\big(E(Q_{\text{target}}), E(Q'_i)\big)$$
$$\text{s.t.} \quad \theta_{\min} \le \text{sim}_{\cos}\big(E(Q_{\text{target}}), E(Q'_i)\big) \le \theta_{\max} \tag{2}$$

Here, $E(\cdot)$ is a sentence-embedding function (e.g., Sentence-BERT (Reimers and Gurevych 2019)), and we set $\theta_{\min} = 0.8$ and $\theta_{\max} = 0.95$ in our experiments. **These thresholds are determined empirically on a validation set, aiming to strike a balance between semantic similarity (relevance) and distinctiveness (contrast).**

Generating Discriminative Representations This stage constitutes the core reasoning step of the RaCoT framework. The triplet $(Q_{\text{target}}, Q_{\text{contrast}}, \Delta)$ is fed into the reasoning model $\mathcal{M}_{\text{RaCoT}}$, which, guided by the contrastive prompting template $\mathcal{T}_{\Delta}$, produces the enhanced query representation Q^*:

$$Q^* = \mathcal{M}_{\text{RaCoT}}\big(\mathcal{T}_{\Delta}(Q_{\text{target}}, Q_{\text{contrast}}, \Delta)\big) \tag{3}$$

Unlike standard retrieval inputs represented as static vectors, Q^* is a semantically enriched object, for instance, a

hypothetical textual sketch of an ideal supporting document, which explicitly encodes strategies to avoid semantic distractors via contrastive reasoning.

RaCoT-Guided Retrieval and Generation

We utilize the enhanced representation Q^* to query the retriever $\mathcal{R}$, obtaining a candidate document set:

$$D_{\text{cand}} = \mathcal{R}(Q^*, \mathcal{C}, K) \tag{4}$$

To improve retrieval quality, we introduce a lightweight *discriminative re-scoring* step. For each document $d_i \in D_{\text{cand}}$, a relevance score is computed as:

$$s_i = p(y = \text{"Relevant"} \mid d_i, Q_{\text{target}}, \Delta) \tag{5}$$

This score is estimated by $\mathcal{M}_{\text{RaCoT}}$ using a classification prompt $\mathcal{T}_{\text{filter}}$. Documents with scores above a threshold τ (set to 0.7 **based on validation set performance**) form the refined context:

$$\mathcal{C}_{\text{final}} = \{d_i \mid d_i \in D_{\text{cand}},\ s_i > \tau\} \tag{6}$$

Finally, the generator $\mathcal{M}_{\text{gen}}$ produces the answer conditioned on both $\mathcal{C}_{\text{final}}$ and the contrastive difference Δ:

$$A = \mathcal{M}_{\text{gen}}(\mathcal{T}_{\text{ans}}(Q_{\text{target}}, \mathcal{C}_{\text{final}}, \Delta)) \tag{7}$$

This ensures that contrastive reasoning is consistently preserved throughout both retrieval and generation stages.

Experiment

Experimental Settings

Evaluation Tasks We evaluate RaCoT on six representative QA benchmarks covering diverse reasoning demands. PopQA (Mallen et al. 2022) and TriviaQA-unfiltered(Joshi et al. 2017) contain a high proportion of long-tail questions involving rare entities and low-frequency knowledge, making them particularly challenging for retrieval-based models due to sparse coverage and limited training signal. ARC-Challenge (Clark et al. 2018) and OpenBookQA (Mihaylov et al. 2018) assess the model's ability to leverage structured scientific and commonsense knowledge. HotpotQA (Yang et al. 2018) and 2WikiMultiHopQA (Ho et al. 2020) require multi-hop reasoning across multiple sentences or documents, testing the model's compositional semantics and ability to track complex inference chains.

Baselines To systematically evaluate the generalization and practical effectiveness of **RaCoT** across different model architectures, we deploy it on two widely-used open-source backbone models: **Qwen2.5-7B** (Team 2024) and **LLaMA3-8B** (AI@Meta 2024), and compare it with a diverse set of representative baselines. We categorize these baselines into three major paradigms: (1) **Privately-enhanced models**, such as *ChatGPT-4o* and *ChatGPT-4o-mini* (OpenAI 2024), which represent the current performance upper bound of commercial closed-source systems; (2) **No-Retrieval methods**, where the model relies solely on its parametric knowledge to answer questions; and (3) **Retrieval-Augmented methods**, which follow the standard retrieve-then-generate paradigm by incorporating external knowledge sources.

Method	PopQA	TQA	ARC-C	OBQA	HOTPOTQA	2WIKI
Proprietary LLM with Retrieval						
GPT-4o	45.3	56.4	53.2	45.6	47.2	36.5
GPT-4o-mini	35.1	60.9	51.8	44.2	45.7	34.2
Baselines without Retrieval						
Qwen2.5-7B	27.8	41.3	41.3	39.4	25.8	31.2
Qwen2.5-7B-Chat	54.6	45.4	62.4	60.9	16.2	28.6
Qwen2.5-7B (SFT)	56.7	59.6	63.9	62.8	43.2	43.2
Baselines with Retrieval						
Qwen2.5-7B	32.4	43.4	43.6	42.6	28.6	32.6
Qwen2.5-7B-Chat	56.5	46.2	65.2	58.3	18.6	30.3
Qwen2.5-7B (SFT)	57.2	61.3	64.2	61.4	44.3	45.6
SAIL-7B	53.2	58.3	59.4	60.4	46.2	48.5
Self-RAG	63.2	65.6	68.4	81.2	65.2	57.4
RQ-RAG	64.5	67.4	69.6	83.9	67.3	58.2
AutoRAG	65.2	68.9	70.3	85.5	67.9	59.1
RankRAG	66.4	70.2	71.4	87.5	68.5	60.3
IterDRAG	66.8	69.8	71.2	86.9	68.6	60.6
Our-RaCoT	**68.3**	**71.8**	**72.1**	**88.2**	**68.9**	**61.2**

Table 2: Benchmark comparison across QA datasets on Qwen models. Our-RaCoT consistently outperforms retrieval baselines and proprietary LLMs.

In each paradigm, we include the base models, their instruction-tuned variants (e.g., *Chat* models), as well as models fine-tuned on specific QA datasets to ensure a fair and comprehensive evaluation. Moreover, we include several recent and high-performing RAG-based methods, such as **SAIL-7B**(Luo et al. 2023), **Self-RAG** (Asai et al. 2024), **RQ-RAG** (Chan et al. 2024), **AutoRAG** (Kim et al. 2024), **RankRAG** (Yu et al. 2024), and **IterDRAG** (Yue et al. 2025), as strong retrieval enhanced baselines. These comparisons allow us to rigorously validate the robustness and effectiveness of RaCoT across various model settings and task scenarios.

Benchmark Model Comparison Experiment

We conduct a comprehensive evaluation of the RaCoT framework across a diverse set of representative QA benchmarks. As shown in Table 2 and Table 3, RaCoT consistently delivers strong and stable performance across different model architectures and task types. On structured reasoning tasks such as ARC-Challenge and OpenBookQA, RaCoT achieves 72.1% / 71.2% on Qwen2.5 and LLaMA3 for ARC-Challenge, and 73.3% / 71.6% for OpenBookQA, matching or slightly outperforming the current best-performing baseline RQ-RAG. These results validate RaCoT's effectiveness in complex reasoning and knowledge integration.

More importantly, RaCoT exhibits notable advantages on long-tail QA benchmarks, which involve rare entities and low-frequency knowledge. PopQA and TriviaQA-unfiltered are particularly challenging datasets for retrieval-augmented methods, as they require robust coverage and fine-grained aggregation of sparse knowledge. On PopQA, RaCoT achieves 68.3% and 59.9% on Qwen2.5 and LLaMA3, re-

Method	PopQA	TQA	ARC-C	OBQA	HOTPOTQA	2WIKI
Baselines without retrieval						
llama3-8B	15.4	30.5	28.8	35.2	6.8	17.2
llama3-8B-chat	41.6	46.2	59.2	56.3	3.2	10.3
llama3-8B (SFT)	46.4	55.3	62.3	54.0	33.6	35.7
Baselines with retrieval						
llama3-8B	18.6	42.5	28.3	37.4	17.2	19.4
llama3-8B-chat	43.5	48.6	58.6	53.2	7.4	11.2
llama3-8B (SFT)	47.9	57.3	58.9	51.3	38.6	37.9
SAIL-7B	42.6	46.2	48.4	51.6	45.6	38.6
Self-RAG	53.6	66.4	67.1	76.4	59.6	43.1
RQ-RAG	54.8	68.9	67.9	79.3	61.8	44.6
AutoRAG	56.3	70.3	68.4	80.5	62.6	45.2
RankRAG	58.5	71.6	69.9	81.7	63.9	46.2
IterDRAG	58.2	72.2	70.3	81.5	64.2	46.4
Our-RaCoT	**59.9**	**73.8**	**71.2**	**81.9**	**65.1**	**47.4**

Table 3: Benchmark comparison across QA datasets on LLaMA3 models. Our-RaCoT outperforms strong retrieval baselines.

spectively, outperforming the best baselines by 1.5–2.4 percentage points. Similarly, on TriviaQA-unfiltered, RaCoT obtains 71.8% and 73.8%, surpassing competing methods by 1.6–1.7 points. These improvements demonstrate RaCoT's enhanced capability in relevance estimation and cross-document semantic integration under knowledge-sparse settings, effectively mitigating the retrieval and reasoning limitations of prior RAG approaches on long-tail distributions.

In summary, RaCoT not only maintains strong performance on mainstream QA tasks but also significantly expands the generalization frontier under long-tail knowledge settings. It exhibits improved robustness in low-frequency entity recognition, semantic evidence aggregation, and multi-document reasoning, thereby addressing key structural weaknesses of existing RAG systems in handling rare and underrepresented knowledge.

Adversarial Distractor Injection

Retrieval Distractor Confusion To further assess the robustness of **RaCoT** against semantically irrelevant but lexically similar distractors, we conduct controlled experiments on two long-tail QA benchmarks: **PopQA** and **TriviaQA-unfiltered**. Specifically, during retrieval, we deliberately inject *distractor passages*, i.e., texts that are lexically similar to the query but semantically irrelevant, into the retrieved context pool.

We compare RaCoT against several strong retrieval-augmented baselines, including **RAG**, **Self-RAG**, **RQ-RAG**, **RankRAG**, and **IterDRAG**, under two complementary metrics: (1) the absolute drop in answer accuracy before and after distractor injection, and (2) the distractor citation rate, defined as the proportion of answers that explicitly rely on distractor content during generation. As shown in Figure 3, RaCoT consistently achieves the smallest performance drop across both datasets (only -8.7% on PopQA and -11.4% on TriviaQA), outperforming the next-best baseline by a large margin. In addition, RaCoT exhibits the lowest distractor citation rate (18% and 24%, respectively), indicating a stronger ability to resist misleading lexical cues. In contrast, other baselines such as Self-RAG and RQ-RAG suffer from substantial degradation (e.g., -19.6% on PopQA and -21.9% on TriviaQA) and exhibit high distractor reliance (up to 57%). These results validate RaCoT's core mechanism in mitigating semantic noise and enhancing factual precision under retrieval ambiguity.

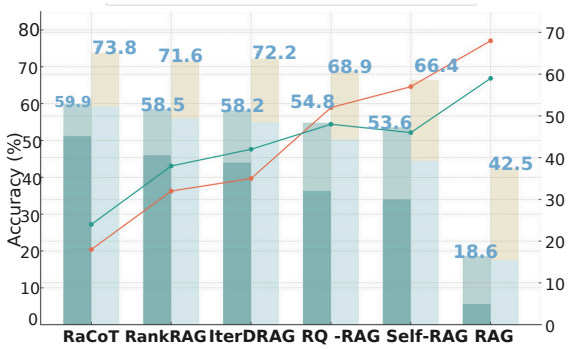

Figure 3: The result of Retrieval Distractor Confusion shows that the green line indicates accuracy and the orange bars indicate distractor citation rate.

Key Issue Identification Interference To further probe the semantic grounding capability of **RaCoT**, we design a controlled experiment to assess its robustness under semantically perturbed contexts. We randomly sample 5,000 instances each from **PopQA** and **ARC-Challenge**, and append contrastive distractor passages to the original queries. These passages are intentionally crafted to be either partially relevant or entirely irrelevant, creating misleading contextual cues that challenge the model's ability to focus on core semantics.

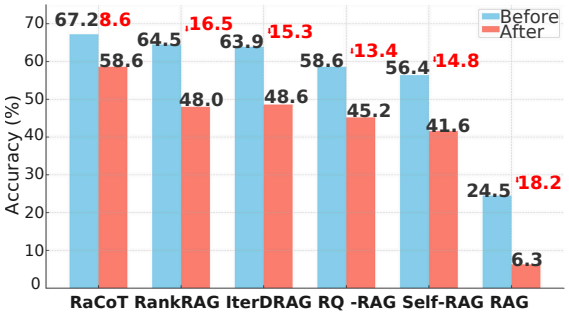

Figure 4: RaCoT$^+$ maintains accuracy under distraction, unlike other RAG methods.

As shown in Figure 4, **RaCoT** achieves a minimal performance degradation of only **8.6%**, significantly outperforming all other retrieval-augmented baselines. In comparison, strong competitors such as RankRAG and IterDRAG suffer drops of **16.5%** and **15.3%**, respectively. Models like Self-RAG and RQ-RAG are even more susceptible, experiencing severe degradation exceeding **20%**. This suggests that

while existing RAG-based systems can exploit lexical overlap, they often lack the semantic discernment needed to filter out irrelevant content.

The superior robustness of RaCoT highlights its key advantage: by contrasting structurally similar yet semantically distinct passages during retrieval-time prompting, RaCoT effectively anchors reasoning on semantically salient cues and avoids overfitting to surface-level signals. These results reaffirm RaCoT's capability to maintain high factual accuracy even under adversarial or misleading information injection.

Efficiency Analysis

To systematically evaluate the retrieval efficiency and computational overhead of **RaCoT**, we record both *retrieval latency* and *token consumption* during inference on the **PopQA** benchmark. For a more comprehensive comparison, we normalize the standard **RAG** framework to a baseline latency (1.0×), and conduct a controlled evaluation across multiple representative RAG-based methods. As shown in

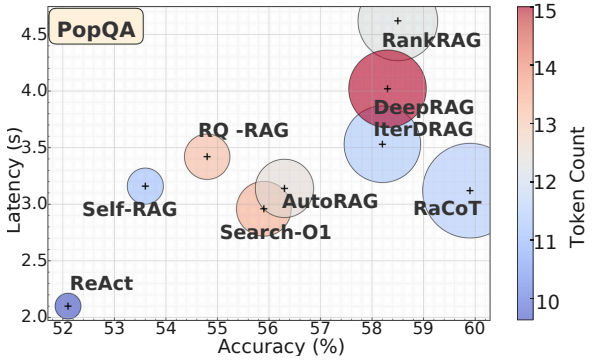

Figure 5: RaCoT+ achieves the highest accuracy with lower latency and token usage. Bubble size denotes answer accuracy, and color reflects average token consumption.

Figure 5, various RAG methods exhibit a clear trade-off among answer accuracy, retrieval latency, and token consumption. Specifically, **RaCoT** achieves the highest accuracy (59.9%) while maintaining relatively low latency (3.12 seconds) and the lowest token usage (11.54), demonstrating dual advantages in both effectiveness and efficiency.

In comparison, methods such as **RankRAG** and **DeepRAG** reach moderately high accuracy but incur substantially higher latency (4.62s and 4.02s) and token usage (12.3 and 15.0) due to large-scale candidate ranking and iterative decoding. Lightweight approaches like **ReAct** achieve the lowest latency (2.1s) and minimal token cost (9.63) but exhibit poor accuracy (52.1%), indicating limited semantic grounding from insufficient retrieval coverage. Methods such as **IterDRAG** and **Search-O1** provide a more balanced cost–quality trade-off, yet remain below the optimal frontier. Overall, these results place **RaCoT** favorably on the accuracy–efficiency Pareto frontier.

Ablation Studies

To evaluate the contribution of each core component in RaCoT, we conduct ablation studies on two long-tail QA benchmarks: PopQA and TQA. The compared variants are as follows: **w/o Contrast Prompting**: Removes the Δ-based prompting, using only the original question to evaluate the benefit of explicit contrastive signals. **w/o Post-Retrieval Ranking**: Skips the reranking step and directly uses the top-5 retrieved documents to assess the impact of reranking on contextual relevance. **w/o Similarity Filtering**: Bypasses the cosine-based filtering during Δ construction and uses the first teacher-generated question to examine the effect of contrastive sample quality. **Weaker Teacher Model**: Replaces the teacher model with a weaker one (e.g., GPT-4o-mini) to assess the model's robustness to teacher strength.

Full RaCoT: The complete system with all components, serving as the performance upper bound.

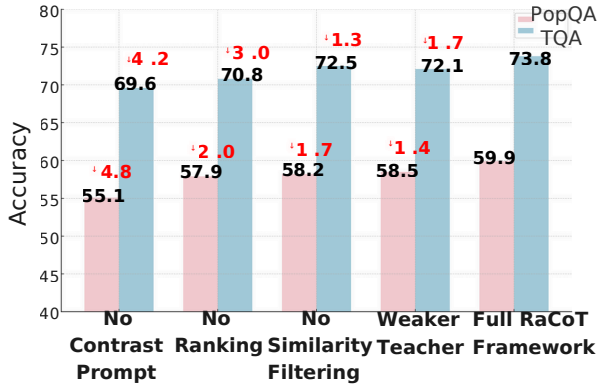

Figure 6: Results of the ablation studies on PopQA and TQA, demonstrating the performance contribution of each component in RaCoT.

As shown in Figure 6, our ablations highlight the complementary roles of all components. Contrastive prompting is the most crucial—its removal yields the largest performance drops (–4.8 on PopQA, –4.2 on TQA), underscoring the importance of explicit difference signals. Post-retrieval reranking and similarity filtering provide additional improvements by refining evidence and maintaining contrast quality, though the system remains strong without them. RaCoT's stability under a weaker teacher further shows that its effectiveness stems from the contrastive mechanism, not teacher strength. Overall, the full RaCoT pipeline reaches top scores (59.9 on PopQA, 73.8 on TQA), demonstrating its robustness for long-tail QA.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMA-JBGS202517, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012985, in part by Guangdong-Hong Kong-Macao Greater Bay Area

Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006, and in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11.

References

AI@Meta. 2024. Llama 3 Model Card.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *ICLR*.

Bommasani, R.; and et al., D. A. H. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.

Brown, T. B.; and et al., B. M. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Chan, C.-M.; Xu, C.; Yuan, R.; Luo, H.; Xue, W.; Guo, Y.; and Fu, J. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Chuang, Y.-S.; Fang, W.; Li, S.-W.; tau Yih, W.; and Glass, J. 2023. Expand, Rerank, and Retrieve: Query Reranking for Open-Domain Question Answering. arXiv:2305.17080.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

de Alvis, C.; and Seneviratne, S. 2024. A Survey of Deep Long-Tail Classification Advancements. arXiv:2404.15593.

Eiter, T.; Geibinger, T.; Higuera, N.; and Oetsch, J. 2023. A logic-based approach to contrastive explainability for neurosymbolic visual question answering. In *IJCAI*.

Fang, Y.; Sun, T.; Shi, Y.; and Gu, X. 2025. AttentionRAG: Attention-Guided Context Pruning in Retrieval-Augmented Generation. arXiv:2503.10720.

Fu, Y.; Peng, H.; Sabharwal, A.; Clark, P.; and Khot, T. 2023. Complexity-Based Prompting for Multi-step Reasoning. In *ICLR*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Grattafiori, A. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Gupta, S.; Ranjan, R.; and Singh, S. N. 2024. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. arXiv:2410.12837.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv:2002.08909.

Ho, X.; Duong Nguyen, A.-K.; Sugawara, S.; and Aizawa, A. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6609–6625. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Huang, Y.; Giledereli, B.; Köksal, A.; Özgür, A.; and Ozkirimli, E. 2021. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *EMNLP*, 8153–8161.

Izacard, G.; and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv:2007.01282.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, arXiv:1705.03551.

Kim, D.; Kim, B.; Han, D.; and Eibich, M. 2024. AutoRAG: Automated Framework for optimization of Retrieval Augmented Generation Pipeline. arXiv:2410.20878.

Lewis, P.; and et al., E. P. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Li, H.; Shi, Y.; Lin, S.; Gu, X.; Lian, H.; Wang, X.; Jia, Y.; Huang, T.; and Wang, Q. 2025a. SWE-Debate: Competitive Multi-Agent Debate for Software Issue Resolution. arXiv:2507.23348.

Li, W.; Hu, B.; Shao, R.; Shen, L.; and Nie, L. 2025b. Lion-fs: Fast & slow video-language thinker as online video assistant. In *CVPR*, 3240–3251.

Li, W.; Zhang, R.; Shao, R.; He, J.; and Nie, L. 2025c. Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification. *arXiv preprint arXiv:2508.21046*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023a. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023b. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.

Liu, W.; Chen, J.; Ji, K.; Zhou, L.; Chen, W.; and Wang, B. 2024. RAG-Instruct: Boosting LLMs with Diverse Retrieval-Augmented Instructions. *arXiv preprint arXiv:2501.00353*.

Liu, W.; Xu, J.; Yu, F.; Lin, Y.; Ji, K.; Chen, W.; Xu, Y.; Wang, Y.; Shang, L.; and Wang, B. 2025. QFFT, Question-Free Fine-Tuning for Adaptive Reasoning. *arXiv preprint arXiv:2506.12860*.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. arXiv:1904.05160.

Luo, H.; Chuang, Y.-S.; Gong, Y.; Zhang, T.; Kim, Y.; Wu, X.; Fox, D.; Meng, H.; and Glass, J. 2023. SAIL: Search-Augmented Instruction Learning. arXiv:2305.15225.

Ma, X.; Gong, Y.; He, P.; Zhao, H.; and Duan, N. 2023. Query Rewriting for Retrieval-Augmented Large Language Models. arXiv:2305.14283.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Hajishirzi, H.; and Khashabi, D. 2022. When Not to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories. *arXiv preprint*.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Qu, Y.; and Ding, Y. e. a. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *ACL*, 5835–5847.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *OpenAI Blog*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*.

Sun, S.; Liu, Y.; Wang, S.; Zhu, C.; and Iyyer, M. 2023. PEARL: Prompting Large Language Models to Plan and Execute Actions Over Long Documents. arXiv:2305.14564.

Tang, J.; Zhang, J.; Lv, Q.; Liu, S.; Yang, J.; Tang, C.; and Wang, K. 2025. HiVA: Self-organized Hierarchical Variable Agent via Goal-driven Semantic-Topological Evolution. arXiv:2509.00189.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Tversky, A. 1977. Features of similarity. *Psychological Review*, 84(4): 327–352.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Wang, X.; Chen, X.; Ren, W.; Han, Z.; Fan, H.; Tang, Y.; and Liu, L. 2024a. Compensation Atmospheric Scattering Model and Two-Branch Network for Single Image Dehazing. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(4): 2880–2896.

Wang, X.; Ren, W.; Chen, X.; Fan, H.; Tang, Y.; and Han, Z. 2024b. Uni-YOLO: Vision-Language Model-Guided YOLO for Robust and Fast Universal Detection in the Open World. In *ACM MM*, MM '24, 1991–2000.

Xiong, W.; and et al., X. L. 2021. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *ICLR*.

Yan, T. L.; Wang, F.; Huang, J. Y.; Zhou, W.; Yin, F.; Galstyan, A.; Yin, W.; and Chen, M. 2024. Contrastive Instruction Tuning. arXiv:2402.11138.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*.

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. arXiv:2407.02485.

Yue, Z.; Zhuang, H.; Bai, A.; Hui, K.; Jagerman, R.; Zeng, H.; Qin, Z.; Wang, D.; Wang, X.; and Bendersky, M. 2025. Inference Scaling for Long-Context Retrieval Augmented Generation. arXiv:2410.04343.

Zhang, J.; Cai, K.; Fan, Y.; Liu, N.; and Wang, K. 2025a. MAT-Agent: Adaptive Multi-Agent Training Optimization. arXiv:2510.17845.

Zhang, J.; Cai, K.; Fan, Y.; Wang, J.; and Wang, K. 2025b. CF-VLM:CounterFactual Vision-Language Fine-tuning. arXiv:2506.17267.

Zhang, J.; Cai, K.; Yang, J.; Wang, J.; Tang, C.; and Wang, K. 2025c. Top-Down Semantic Refinement for Image Captioning. arXiv:2510.22391.

Zhang, J.; Cai, K.; Yang, J.; and Wang, K. 2025d. Learning Dynamics of VLM Finetuning. arXiv:2510.11978.

Zhang, J.; Cai, K.; Zeng, Q.; Liu, N.; Fan, S.; Chen, Z.; and Wang, K. 2025e. Failure-Driven Workflow Refinement. arXiv:2510.10035.

Zhang, J.; Fan, Y.; Cai, K.; Huang, Z.; Sun, X.; Wang, J.; Tang, C.; and Wang, K. 2025f. DrDiff: Dynamic Routing Diffusion with Hierarchical Attention for Breaking the Efficiency-Quality Trade-off. arXiv:2509.02785.

Zhang, J.; Fan, Y.; Cai, K.; Sun, X.; and Wang, K. 2025g. OSC: Cognitive Orchestration through Dynamic Knowledge Alignment in Multi-Agent LLM Collaboration. arXiv:2509.04876.

Zhang, J.; Fan, Y.; Lin, W.; Chen, R.; Jiang, H.; Chai, W.; Wang, J.; and Wang, K. 2025h. GAM-Agent: Game-Theoretic and Uncertainty-Aware Collaboration for Complex Visual Reasoning. arXiv:2505.23399.

Zhang, J.; Huang, Z.; Fan, Y.; Liu, N.; Li, M.; Yang, Z.; Yao, J.; Wang, J.; and Wang, K. 2025i. KABB: Knowledge-Aware Bayesian Bandits for Dynamic Expert Coordination in Multi-Agent Systems. In *ICML*.

Zhang, N.; Deng, S.; Sun, Z.; Wang, G.; Chen, X.; Zhang, W.; and Chen, H. 2019. Long-tail Relation Extraction via Knowledge Graph Embeddings and Graph Convolution Networks. arXiv:1903.01306.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023. Deep Long-Tailed Learning: A Survey. *T-PAMI*, 45(9): 10795–10816.

Zhao, W. X.; and et al., K. Z. 2025. A Survey of Large Language Models. arXiv:2303.18223.

Zhou, K.; and Gong, Y. e. a. 2022. SimANS: Simple Ambiguous Negatives Sampling for Dense Text Retrieval. In Li, Y.; and Lazaridou, A., eds., *EMNLP*, 548–559.