

Guaranteeing Knowledge Integration with Joint Decoding for Retrieval-Augmented Generation

Zhengyi Zhao^{1,5}, Shubo Zhang², Zezhong Wang^{1,5}, Yuxi Zhang², Huimin Wang³,
Yutian Zhao³, Yefeng Zheng⁴, Binyang Li^{2*}, Kam-Fai Wong^{1,5}, Xian Wu^{3*}

¹ The Chinese University of Hong Kong ² University of International Relations

³ Tencent Jarvis Lab ⁴ Westlake University

⁵ Ministry of Education Key Laboratory of High Confidence Software Technologies, CUHK
{zyzhao, kfwong}@se.cuhk.edu.hk, byli@uir.edu.cn, kevinxwu@tencent.com

Abstract

Retrieval-Augmented Generation (RAG) significantly enhances Large Language Models (LLMs) by providing access to external knowledge. However, current research primarily focuses on retrieval quality, often overlooking the critical “integration bottleneck”: even when relevant documents are retrieved, LLMs frequently fail to utilize them effectively due to conflicts with their internal parametric knowledge. In this paper, we argue that implicitly resolving this conflict in a single generation pass is suboptimal. We introduce GUARANTRAG, a framework that explicitly decouples reasoning from evidence integration. First, we generate an “Inner-Answer” based solely on parametric knowledge to capture the model’s reasoning flow. Second, to guarantee faithful evidence extraction, we generate a “Refer-Answer” using a novel Contrastive DPO objective. This objective treats the parametric Inner-Answer as a negative constraint and the retrieved documents as positive ground truth, forcing the model to suppress internal hallucinations in favor of external evidence during this phase. Finally, rather than naive concatenation or using the DPO trained model directly, we propose a joint decoding mechanism that dynamically fuses the logical coherence of the Inner-Answer with the factual precision of the Refer-Answer at the token level. Experiments on five QA benchmarks demonstrate that GUARANTRAG improves accuracy by up to 12.1% and reduces hallucinations by 16.3% compared to standard and dynamic RAG baselines.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities but often struggle with factual accuracy in knowledge-intensive tasks (Jiang et al., 2023a; Zeng et al., 2024; Xiong et al., 2024). Retrieval-Augmented Generation (RAG) addresses this by supplementing

the model’s parametric memory with external documents (Dong et al., 2025; Lin et al., 2025). While recent advancements have optimized retrieval precision and dynamic triggering (Wu et al., 2024; Jeong et al., 2024; Yang et al., 2024), a critical challenge remains: the effective *integration* of retrieved knowledge.

Current RAG methods typically feed retrieved documents and user queries into the LLM, expecting it to implicitly reconcile the external information with its internal knowledge. However, this approach often leads to two failure modes: (1) *Parametric Override*, where the model ignores retrieved details (e.g., specific statistics) in favor of its generalized internal priors, and (2) *Disjointed Integration*, where external facts are inserted clumsily, disrupting the reasoning flow. Our preliminary experiments on 500 knowledge-intensive queries reveal that 67.3% of RAG responses suffer from such integration failures, suggesting that the conflict between parametric and non-parametric knowledge is a fundamental bottleneck that cannot be solved by retrieval improvements alone.

To bridge this gap, we introduce GUARANTRAG, a framework that transforms knowledge integration from an implicit “black box” process into an explicit, multi-stage pipeline. Our core insight is to divide the generation task into two distinct objectives: maintaining reasoning coherence (Parametric) and ensuring factual faithfulness (Non-Parametric).

GUARANTRAG operates in three phases. First, we generate an “Inner-Answer” using the model’s internal knowledge. This captures the LLM’s superior reasoning structure and linguistic fluency, albeit with potential hallucinations. Second, we generate a “Refer-Answer” dedicated strictly to evidence extraction. A key innovation here is our training strategy: we observe that standard supervised fine-tuning often fails to suppress the model’s strong internal priors. To address this, we employ

*Corresponding Author

Direct Preference Optimization (DPO) specifically for the Refer-Answer generation. By treating the retrieved documents as the “chosen” source and the model’s own Inner-Answer as the “rejected” negative sample, we explicitly penalize the model for relying on parametric memory when it conflicts with external evidence. This ensures the Refer-Answer is a faithful distillation of the retrieved content. More importantly, we are now aiming to adopt the DPO trained model directly to have the final answer. The model here is only to explicitly distinguish the answer with and without retrieved documents.

Finally, we address the challenge of fusing these two distinct outputs. Simple concatenation of the Inner and Refer answers is insufficient, as it often leads to redundancy or exceeds context windows, causing the model to revert to its priors due to attention dilution. Instead, we propose a Joint Decoding mechanism. This method operates at the token generation level, using the high-level reasoning skeleton of the Inner-Answer while dynamically substituting factual tokens from the Refer-Answer when semantic divergence is detected. This allows GUARANTRAG to combine the “best of both worlds”: the coherent logic of the LLM and the strict accuracy of the retrieval system. Our contributions are as follows:

- We identify the conflict between parametric and non-parametric knowledge as the primary cause of RAG integration failure and propose GUARANTRAG to explicitly decouple and refuse these knowledge sources.
- We introduce a Contrastive DPO training objective that utilizes the model’s own Inner-Answer as a negative constraint, guaranteeing that the Refer-Answer is grounded in retrieved evidence rather than internal memory.
- We develop a Joint Decoding mechanism that synergizes the reasoning flow of internal knowledge with the factual precision of external documents, outperforming naive concatenation strategies. And extensive experiments show that GUARANTRAG improves answer accuracy by up to 12.1% and reduces hallucinations by 16.3% across five standard benchmarks.

2 Related Works

RAG for QA. RAG has significantly advanced QA capabilities through three primary approaches: (1) static RAG methods (Jiang et al., 2023a,

2024; Laitenberger et al., 2025) that apply a fixed retrieval-then-generate pipeline; (2) adaptive RAG approaches (Siriwardhana et al., 2023; Wu et al., 2024; Jeong et al., 2024) that dynamically determine when to retrieve based on query characteristics; and (3) iterative RAG systems (Wang et al., 2024; Macdonald et al., 2025; Hayashi et al., 2025) that refine outputs through multiple retrieval-generation cycles. Despite these advances, existing methods still struggle with effectively integrating parametric and non-parametric knowledge. While recent works have improved retrieval precision (Kalra et al., 2024; Rezaei and Dieng, 2025) and query reformulation (Wang et al., 2024; Hayashi et al., 2025), they typically treat retrieval and generation as sequential rather than interacting processes.

Answer Fusion Techniques. Answer fusion research has explored three primary integration strategies: (1) context-level fusion (Sun et al., 2018; Deng et al., 2020; Wang et al., 2023), which combines retrieved documents with queries before generation but often leads to information overload; (2) token-level fusion (Chen et al., 2022; Mordo et al., 2024), which integrates information during token generation but frequently sacrifices narrative coherence; and (3) answer-level fusion (Rackauckas, 2024; Rackauckas et al., 2024; Sivasothy et al., 2024), which combines separately generated responses but typically operates at coarse granularity. While recent advances in attention mechanisms (Li et al., 2025; Fang et al., 2025) and ensemble methods (Gan et al., 2025; Chen et al., 2025) have improved information integration, they still treat retrieved content uniformly without considering its relationship to the model’s parametric knowledge. Our GUARANTRAG framework introduces a novel segment-level contrastive fusion mechanism that operates at a finer granularity than existing methods, explicitly decomposing answers into semantic segments and performing targeted alignment between complementary knowledge sources.

3 Methodology

We propose GUARANTRAG, a framework designed to resolve the conflict between parametric memory and external evidence in RAG systems. As illustrated in Figure 1, our approach operates in three stages: (1) a *Knowledge Decision* module that filters unnecessary retrieval; (2) a *Dual-Path Generation* phase that explicitly decouples the model’s

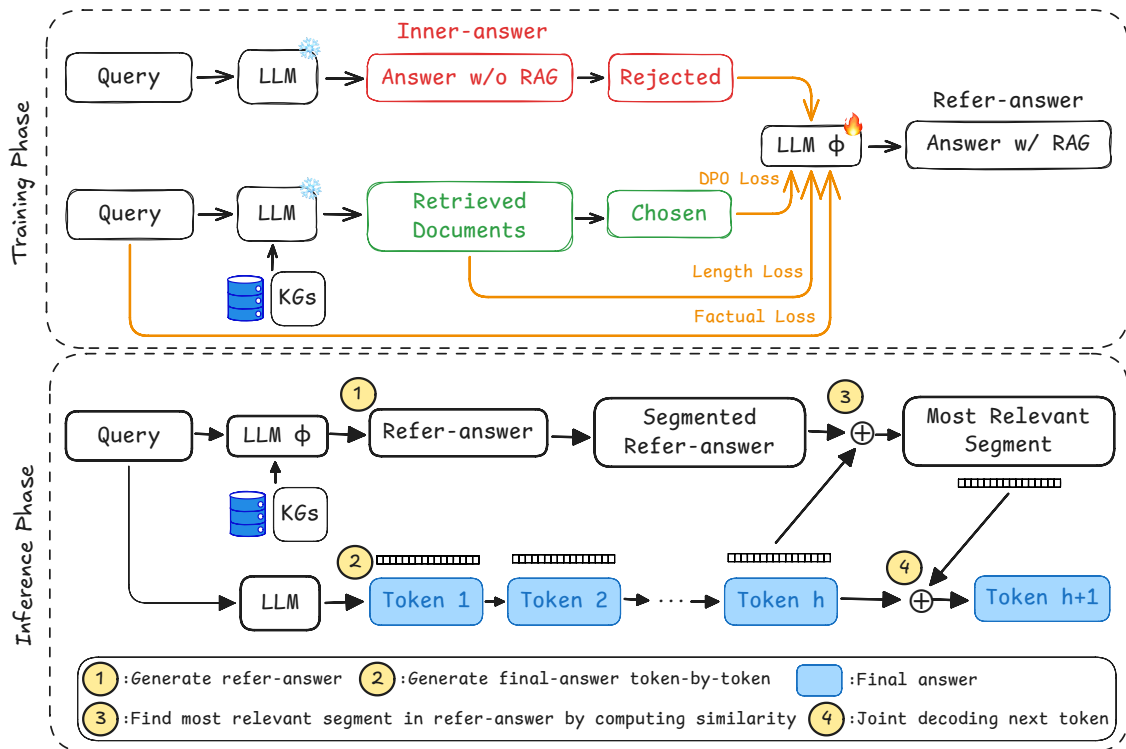


Figure 1: Overview of the GUARANTRAG framework. (1) **Decoupling**: We generate an Inner-Answer (reasoning) and a Refer-Answer (evidence). The Refer-Answer is trained via Contrastive DPO to explicitly prefer retrieved docs over parametric priors. (2) **Fusion**: A Joint Decoding mechanism dynamically merges the reasoning flow of the Inner-Answer with the factual content of the Refer-Answer during inference.

reasoning flow (Inner-Answer) from factual evidence (Refer-Answer) via contrastive alignment; and (3) a *Joint Decoding* mechanism that dynamically fuses these streams at the token level.

3.1 Problem Formulation

Given a knowledge-intensive query q and a set of retrieved documents $D = \{d_1, \dots, d_k\}$, the goal of a RAG system is to generate a response y . Standard methods model this as $P(y|q, D)$ directly. However, we posit that this formulation forces the LLM to implicitly resolve the conflict between its pre-trained parametric knowledge $\mathcal{K}_{\text{param}}$ and the non-parametric context D , often resulting in hallucinations or disjointed integration.

We reformulate the generation process by decomposing the output into two intermediate latent states: a reasoning-focused *Inner-Answer* y_{inner} derived from $\mathcal{K}_{\text{param}}$, and an evidence-focused *Refer-Answer* y_{ref} derived from D . The final response y is generated via a joint decoding policy π_{joint} that harmonizes these states:

$$y \sim \pi_{\text{joint}}(\cdot | q, y_{\text{inner}}, y_{\text{ref}}) \quad (1)$$

3.2 Phase 1: Knowledge Decision

To ensure computational efficiency, we first determine if retrieval is necessary. We fine-tune a lightweight estimator based on the base LLM to predict a binary retrieval signal $z \in \{0, 1\}$. The estimator evaluates: (1) **Temporal Relevance**: is the query outside the training cutoff? and (2) **Factual Specificity**: does the query require long-tail entity knowledge? If $z = 0$, we generate directly. If $z = 1$, we proceed to the dual-path framework.

3.3 Phase 2: Dual-Path Answer Generation

This phase generates the two intermediate states required for our joint decoding.

Path A: Inner-Answer Generation. We first utilize the base model π_{θ} to generate an inner-answer $y_{\text{inner}} \sim \pi_{\theta}(\cdot | q)$. While potentially factually inaccurate, y_{inner} preserves the model’s optimal reasoning structure, linguistic fluency, and instruction-following capabilities. It serves as a “structural template” for the final generation.

Path B: Refer-Answer Generation. The core challenge in RAG is ensuring the model prioritizes

D over $\mathcal{K}_{\text{param}}$. To address this, we train a specialized Evidence Model π_ϕ , which initialized from π_θ , to generate a *Refer-Answer* y_{ref} . Unlike standard fine-tuning, we employ a Contrastive DPO objective to explicitly suppress parametric hallucinations. We construct preference pairs where the retrieved document content D is the *chosen* response y_w , and the model’s own hallucinated y_{inner} is the *rejected* response y_l . The objective is:

$$\mathcal{L}_{\text{DPO}}(\phi) = -\log \sigma \left(\beta \log \frac{\pi_\phi(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\phi(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \quad (2)$$

where $x = (q, D)$. This forces π_ϕ to treat parametric priors as negative constraints, ensuring y_{ref} is a faithful reflection of external evidence. **It should be noted that although we regard the inner-answer as reject sample, sometimes the inner-answer is already the correct answer. The purpose here is to use DPO training the model to distinguish the two kinds of answer, instead of adopt this model directly to have the final answer.**

To prevent the degeneration often seen in DPO (e.g., verbatim copying or loss of coherence), we incorporate two auxiliary constraints: 1. *Adaptive Length Regularization*: We align the length of the evidence summary with the reasoning path to ensure compatibility during fusion:

$$\mathcal{L}_{\text{len}}(\phi) = \max(0, |y_{\text{ref}}| - |y_{\text{inner}}|)^2 \quad (3)$$

And 2. *Factual Relevance Constraint*: To ensure semantic alignment with the query, we maximize the cosine similarity between the embeddings of the query $E(q)$ and the generated refer-answer $E(y_{\text{ref}})$:

$$\mathcal{L}_{\text{fact}}(\phi) = 1 - \cos(E(q), E(y_{\text{ref}})) \quad (4)$$

The final training objective for the Evidence Model is $\mathcal{L} = \mathcal{L}_{\text{DPO}} + \lambda_1 \mathcal{L}_{\text{len}} + \lambda_2 \mathcal{L}_{\text{fact}}$.

3.4 Phase 3: Joint Decoding with Dynamic Fusion

The final step is to synthesize the reasoning of y_{inner} with the facts of y_{ref} . Naive concatenation of these outputs often fails due to attention dilution. Instead, we propose a Dynamic Joint Decoding mechanism that operates at the hidden-state level.

LLM-based Semantic Segmentation. To facilitate precise knowledge injection, we first decompose the evidence-rich refer-answer y_{ref} into a set of discrete semantic segments $S_{\text{ref}} = \{s_1, \dots, s_m\}$. Unlike heuristic splitting (e.g., by punctuation) which often leaves multiple distinct facts entangled within complex sentences, we employ a lightweight auxiliary LLM to break the text into *atomic semantic units*. This LLM-driven decomposition is critical for two reasons: (1) **Granularity**: It ensures that each segment s_j contains exactly one piece of factual evidence, preventing the model from retrieving irrelevant context during fusion; and (2) **Robustness**: It resolves ambiguous pronoun references within sentences, ensuring that segments remain semantically self-contained for accurate similarity matching during the decoding phase.

Token-Level Fusion. We generate the final response y using the base model π_θ . At each decoding step t , let $\mathbf{h}_t^{\text{gen}}$ denote the hidden state of the current token being generated. We simultaneously maintain a context window of the currently generated sentence segment, denoted as s_{curr} . We compute the relevance of the current generation context to the evidence segments:

$$w_j = \text{softmax} \left(\frac{\mathbf{h}(s_{\text{curr}})^\top \mathbf{h}(s_j)}{\tau} \right) \quad (5)$$

where $\mathbf{h}(\cdot)$ is a sentence encoder. We identify the most relevant evidence segment $s^* = s_{\arg \max(w_j)}$.

To integrate this evidence, we perform a *soft intervention* on the hidden state before the projection to the vocabulary space. The enhanced hidden state $\tilde{\mathbf{h}}_t$ is computed as:

$$\tilde{\mathbf{h}}_t = \mathbf{h}_t^{\text{gen}} + \gamma \cdot \mathbf{h}(s^*) \quad (6)$$

where γ is a gating factor controlling the strength of external knowledge injection. The final token probability is computed via $P(y_t|y_{<t}) = \text{Softmax}(W_v \tilde{\mathbf{h}}_t)$. This mechanism allows the model to follow the reasoning skeleton of its parametric knowledge $\mathbf{h}_t^{\text{gen}}$ while being dynamically “steered” toward factual accuracy by the retrieved evidence $\mathbf{h}(s^*)$ whenever the context aligns.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate on five diverse QA benchmarks: Natural Questions (NQ) (Kwiatkowski

et al., 2019), TruthfulQA (Lin et al., 2022), Wizard of Wikipedia (WoW) (Dinan et al.), HotpotQA (Yang et al., 2018), and ELI5 (Fan et al., 2019). These datasets span various knowledge integration scenarios from factoid queries to multi-hop reasoning tasks.

Baselines. We compare GUARANTRAG against both non-RAG and state-of-the-art RAG approaches. The non-RAG baseline uses Qwen3-8B without retrieval augmentation. RAG baselines include Standard RAG, SelfRAG (Asai et al., 2023), RQ-RAG (Chan et al., 2024), SOLAR (Kim et al., 2024), DA-RAG (Su et al., 2025), FLARE (Jiang et al., 2023b), DRAGIN (Su et al., 2024), and P-RAG (Su et al., 2025). All methods are implemented with Qwen3-8B for fair comparison. For retrieval, we employ BM25 (Robertson and Walker, 1994), RetroMAE (Xiao et al., 2022), SPLADE-v3 (Lassance et al., 2024), and HyDE (Gao et al., 2023) as different retrievers.

Evaluation Metrics. We evaluate our approach using metrics across three distinct levels of assessment. At the lexicon level, we employ **Match**, **ROUGE-L**, and **BLEU-4**. At the semantic level, we utilize **BERTScore** and **BEM**. For reference-document usage, we introduce three specialized metrics: **Hallucination Rate (Hal.)** to quantify factually unsupported claims, **Entity Precision (Ent.)** to measure accurate entity reproduction from references, and **Structure Coherence (Struc.)** to evaluate logical organization on a 5-point scale through human assessment. Detailed settings and metrics can be found in Appendix B.

4.2 Experimental Results

Table 1 and Table 2 demonstrate that GUARANTRAG consistently achieves superior performance compared to both conventional and state-of-the-art dynamic RAG approaches. The results reveal several critical insights about knowledge integration in RAG systems. First, we observe that the performance gap between GUARANTRAG and baseline methods increases with model capacity, suggesting that our framework effectively leverages stronger parametric knowledge while mitigating integration conflicts. Second, our approach demonstrates remarkable consistency across diverse retrieval mechanisms, indicating that the benefits stem from improved knowledge fusion rather than retrieval optimization. Most importantly, GUARANTRAG successfully resolves the fundamental

trade-off between factual accuracy and response coherence that has plagued existing RAG systems, while achieving the lowest hallucination rates and highest entity precision, our method simultaneously maintains structural coherence scores comparable to vanilla models without retrieval augmentation. This breakthrough suggests that explicit modeling of parametric-nonparametric knowledge conflicts through complementary answer generation and segment-level fusion represents a more principled approach to knowledge integration than existing implicit fusion strategies. Detailed experimental results can be found in Appendix C.1.

4.3 Ablation and Analysis Studies

To understand the effectiveness of each component in GUARANTRAG and analyze its behavior under different conditions, we conduct comprehensive ablation studies and analyses.

Ablation Studies. Table 3 presents a systematic ablation study evaluating the contribution of each core component in GUARANTRAG. The analysis reveals that length control mechanisms provide the most substantial contribution, highlighting the critical importance of constraining refer-answer verbosity while maintaining factual density. Relevance control follows as the second most impactful component, demonstrating that selective filtering of retrieved information based on query-document similarity is essential for effective knowledge integration. Both DPO training and segment-level fusion contribute comparably to performance, indicating that our contrastive learning strategy for refer-answer optimization and the strategic combination mechanism for dual answers are equally vital. To investigate whether our proposed segment-level fusion is exclusively dependent on the DPO-based framework or generally applicable, we conducted a decoupling ablation study on the HotpotQA dataset using the Qwen3-14B backbone. As detailed in Table 4, applying segment-level fusion to a standard baseline model and our method can both show improvements, confirming its standalone value across general RAG architectures.

Segment-level Decomposition obtains best refer-answer matching. To validate the effectiveness of our segment-level fusion mechanism, we conducted an analysis comparing three decomposition granularities (token-level, sentence-level, and segment-level) on 500 randomly sampled query-response pairs, evaluating attention distribution. As

Method	Performance Across Datasets					Average	Reference Usage		
	NQ	TruthfulQA	WoW	HotpotQA	ELI5		Hal.↓	Ent.↑	Struc.↑
<i>Qwen3-8B Based Methods</i>									
Qwen3-8B	59.8	52.3	56.2	48.7	57.1	54.8	38.2	–	4.72
w/ Standard RAG	63.4	58.1	62.7	54.3	61.5	60.0	32.6	76.8	4.15
w/ SelfRAG	67.3	61.4	66.8	59.2	67.4	64.4	28.1	81.3	4.28
w/ RQ-RAG	68.9	62.7	68.1	61.5	68.7	65.9	26.4	82.9	4.31
w/ SOLAR	70.4	64.2	69.5	63.1	70.2	67.5	24.7	84.5	4.37
w/ DA-RAG	66.1	60.8	65.4	57.9	66.2	63.3	29.5	80.1	4.22
w/ FLARE	69.2	63.5	68.7	62.4	69.8	66.7	25.8	83.7	4.35
w/ DRAGIN	69.6	63.9	69.1	62.8	70.1	67.1	25.3	84.2	4.38
w/ P-RAG	71.8	65.4	71.2	64.7	72.1	69.0	22.9	86.1	4.43
w/ GuarantRAG	76.4	69.8	75.1	68.9	74.6	74.0	15.7	88.4	4.68
<i>Qwen3-14B Based Methods</i>									
Qwen3-14B	62.4	54.9	58.7	51.3	59.2	57.3	35.8	–	4.76
w/ Standard RAG	66.8	60.7	65.1	57.2	64.3	62.8	30.4	78.9	4.19
w/ SelfRAG	70.1	64.2	69.3	62.5	70.1	67.2	25.9	83.7	4.32
w/ RQ-RAG	71.9	65.8	70.9	64.3	71.8	68.9	24.1	85.2	4.35
w/ SOLAR	73.6	67.1	72.4	66.0	73.2	70.5	22.3	86.8	4.41
w/ DA-RAG	69.3	63.5	68.2	60.8	69.0	66.2	27.2	82.4	4.26
w/ FLARE	72.4	66.3	71.5	65.1	72.7	69.6	23.5	86.1	4.39
w/ DRAGIN	72.8	66.7	72.0	65.6	73.1	70.0	23.0	86.5	4.42
w/ P-RAG	75.2	68.9	74.3	68.1	75.8	72.5	20.4	88.7	4.47
w/ GuarantRAG	79.1	72.4	77.8	71.6	78.3	75.8	13.2	90.9	4.72
<i>Qwen3-14B Thinking Based Methods</i>									
Qwen3-14B-T	64.1	57.8	59.3	58.7	57.4	59.5	32.9	–	4.79
w/ Standard RAG	68.2	62.1	64.7	63.4	61.8	64.0	34.7	77.2	4.21
w/ SelfRAG	71.4	66.5	68.9	69.1	67.2	68.6	31.2	81.9	4.34
w/ RQ-RAG	73.1	68.2	70.4	70.8	68.9	70.3	29.3	83.4	4.37
w/ SOLAR	74.9	69.7	72.1	72.5	70.6	71.9	27.1	85.1	4.43
w/ DA-RAG	70.6	65.8	67.8	67.3	66.1	67.5	32.5	80.7	4.28
w/ FLARE	73.8	68.9	71.2	71.7	69.8	71.1	28.7	84.3	4.41
w/ DRAGIN	74.2	69.3	71.6	72.1	70.2	71.5	28.2	84.7	4.44
w/ P-RAG	76.4	71.6	73.8	74.9	72.5	73.8	25.6	86.9	4.49
w/ GuarantRAG	80.3	74.1	76.2	78.2	74.8	76.7	18.4	89.3	4.74
GPT-4o	78.3	71.8	77.2	70.9	76.9	75.0	14.1	–	4.81
o1	79.5	73.1	78.6	72.3	78.4	76.4	12.8	–	4.85

Hal.: Hallucination Rate (%); Ent.: Entity Precision (%); Struc.: Structure Coherence (5-point scale)

Table 1: Comprehensive performance comparison across five datasets and knowledge integration quality metrics with BM25 retriever. Results are average five metrics. Reference usage metrics evaluate hallucination mitigation and response coherence. -T denotes thinking mode. **Bold** indicates the best overall performance within each model family.

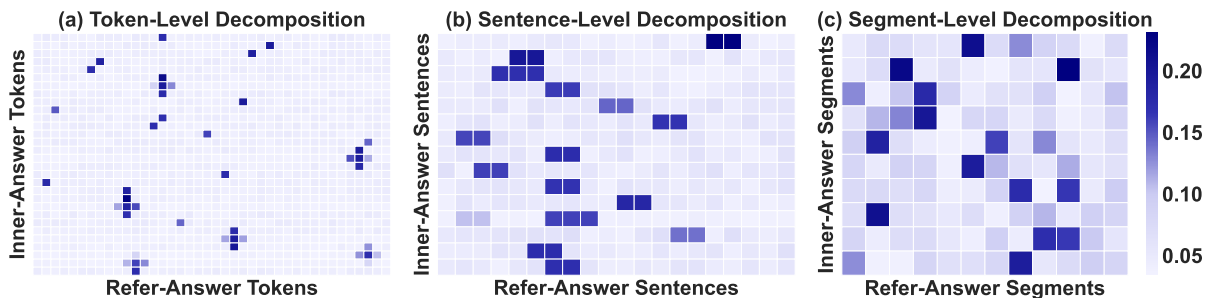


Figure 2: Attention distribution heatmaps for different decomposition granularities.

shown in Figure 2, the token-level approach demonstrates substantial inequality in attention distribution, with Figure 2(a) revealing highly concentrated attention on few tokens while most content receives

minimal attention. The sentence-level approach shows moderate improvement, but still exhibits blocky attention patterns where entire sentences are either significantly attended to or largely ignored

Method	BM25	SPLADE	RetroMAE	HyDE
Qwen3-14B-T	59.5	59.5	59.5	59.5
w/ RAG	62.1	64.3	65.7	64.0
w/ SelfRAG	66.2	69.1	70.6	68.5
w/ RQ-RAG	67.9	70.8	72.4	70.1
w/ SOLAR	69.4	72.5	74.1	71.6
w/ DA-RAG	65.1	68.2	69.4	67.3
w/ FLARE	68.7	71.6	73.1	70.9
w/ DRAGIN	69.1	72.0	73.6	71.2
w/ P-RAG	71.3	74.4	75.8	73.7
GuarantRAG	74.2	77.1	78.9	76.6

Table 2: Performance comparison across different retrievers for Qwen3-14B thinking mode, averaged over all five datasets and five metrics.

Model Variant	Avg.	Δ
GUARANTRAG	74.2	-
w/o DPO Training	70.5	-3.7
w/o Length Control	66.8	-7.4
w/o Relevance Control	69.1	-5.1
w/o Segment-Level Fusion	68.9	-5.3
GUARANTRAG-7B	71.6	-2.6
GUARANTRAG-4B	67.3	-6.9
Qwen3-14B+RAG	62.1	-12.1

Table 3: Ablation study showing the contribution of each component in GUARANTRAG-14B. Results are averaged across all metrics using the BM25 retriever. The lower section shows performance of smaller GuarantRAG variants and the standard RAG baseline for comparison.

(Figure 2(b)). Our segment-level approach significantly outperforms both alternatives across all metrics, with Figure 2(c) demonstrating a more uniform attention distribution where attention meaningfully spans a wider range of semantic segments, enabling more comprehensive knowledge integration by preventing attention concentration on narrow portions of the retrieved information.

GUARANTRAG achieves robust performance across various input complexity. To further understand the performance of GUARANTRAG under varying conditions, we conduct a systematic analysis across different query lengths, reasoning complexities, and reference document lengths. Figure 3 presents this analysis using the RetroMAE retriever. We categorize queries into short (<10 words), medium (10-25 words), and long (>25 words); reasoning complexity as simple, moderate, and complex; and document lengths as concise (<300 tokens), moderate (300-800 tokens), and extensive (>800 tokens)¹. Our results reveal that

¹Detailed settings can be found in Appendix C.2.

GUARANTRAG maintains its performance advantage across all settings, with particularly significant improvements for complex reasoning queries and when dealing with extensive reference documents. The improvements come from our selective joint decoding mechanism which select the most relevant and useful segments to help LLM incorporate the retrieved informations to final answer.

GUARANTRAG has fewer impact on noisy retrieval. To evaluate our system’s impacts on noisy retrieval results, we conducted a controlled experiment by deliberately introducing irrelevant documents into the retrieval set. Specifically, we gradually added 1 to 5 noisy documents to the top-5 relevant documents, resulting in progressively diluted retrieval sets containing 16.7% to 50% irrelevant information. As shown in Figure 4, GUARANTRAG demonstrates superior robustness to retrieval noise compared to baseline approaches. While SelfRAG, RQ-RAG, and SOLAR experience substantial performance degradation, GUARANTRAG maintains 90.3% of its original performance under the same conditions.

Inner-answer maintains structural validity independent of factual grounding. A potential concern with GUARANTRAG’s decoupling strategy is that in complex reasoning tasks, logic often depends heavily on the underlying facts. Consequently, if the model lacks factual knowledge, its parametric inner-answer might exhibit fundamentally flawed logic, risking incoherent or “Frankenstein” responses when fused with external facts. To empirically validate the safety of decoupling reasoning from facts, we evaluated the structural integrity of the inner-answer prior to any RAG integration. Using a random sample of 300 instances from the HotpotQA dataset, we assessed the *Grammatical Coherence* and *Structural Validity* of the generated reasoning skeletons utilizing both human annotators and an LLM-as-a-judge paradigm. As detailed in Table 6, the inner-answers exhibit high structural validity and grammatical coherence, demonstrating that the base model rarely hallucinates fundamentally flawed logic even when facts are missing. This confirms that our decoupled structural templates provide a robust and reliable foundation for subsequent factual fusion.

4.3.1 Computational Efficiency Analysis

Table 5 presents detailed computational overhead analysis across different model sizes and methods,

Architecture Setting	DPO Training	Fusion Strategy	Exact Match (EM)
Vanilla RAG	None	Standard Autoregressive	34.5
Standalone Fusion	None	Segment-Level Fusion	37.1 (+2.6)
GUARANTRAG (DPO Only)	Yes (Ours)	Standard Autoregressive	38.2 (+3.7)
Full GUARANTRAG	Yes (Ours)	Segment-Level Fusion	40.7 (+6.2)

Table 4: Ablation on the standalone effectiveness of Segment-Level Fusion.

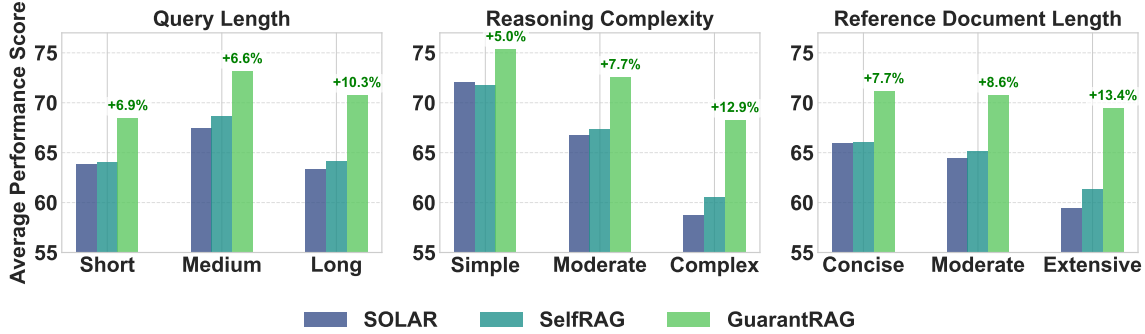


Figure 3: Performance analysis of GUARANTRAG compared to SOLAR and SelfRAG across different (a) query lengths, (b) reasoning complexities, and (c) reference document lengths. Y-axis shows the average performance score across all metrics.

Method	Latency (s)	Reasoning Tokens	Answering Tokens	Total Tokens	Costs	Quality Gain
Standard RAG	3.18	1,847	156	2,003	1.0×	-
SelfRAG	5.41	2,394	203	2,597	1.7×	+7.8%
SOLAR	6.23	2,681	228	2,909	2.0×	+10.1%
P-RAG	5.74	2,543	189	2,732	1.8×	+12.4%
GuarantRAG	4.23	2,156	167	2,323	1.3×	+16.3%

Table 5: Computational efficiency analysis focusing on inference time and token consumption based on Qwen3-14B-Thinking.

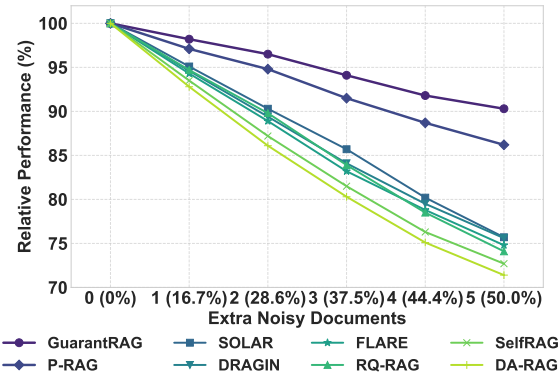


Figure 4: Performance degradation with increasing retrieval noise. x-axis denotes the extra noisy documents added to the model.

focusing on inference time and token consumption. GUARANTRAG demonstrates superior efficiency compared to other advanced RAG methods when considering both inference time and token usage. While requiring 16% more tokens than standard RAG, it introduces only 33% time overhead

Evaluation Metric	Evaluator	Accuracy
Grammatical Coherence	Human	96.8%
Grammatical Coherence	Gemini 3.1 Pro	91.8%
Structural Validity	Human	89.4%
Structural Validity	Gemini 3.1 Pro	93.2%

Table 6: Evaluation of the inner-answer’s structural validity and grammatical coherence prior to RAG integration, based on 300 sampled instances from HotpotQA.

compared to methods like SelfRAG and SOLAR. This efficient token utilization, combined with streamlined processing, enables GUARANTRAG to achieve substantially higher quality gains while maintaining better computational efficiency than competing advanced RAG approaches. We also analyse impacts of different answer fusion mechanisms in Appendix C.3, impacts of different decision making mechanisms in Appendix C.4, and impacts of thinking mode in Appendix C.5.

Query: What was the specific impact of the 2023 earthquake on Turkey’s GDP growth rate?

Retrieved Document: The February 2023 earthquakes in Turkey caused an estimated \$84.1 billion in damages, reducing the country’s GDP growth rate from the projected 5.2% to 2.8% for 2023, according to the World Bank’s post-disaster assessment report.

Standard RAG: The 2023 earthquake in Turkey had a significant economic impact on the country. According to reports (\$84.1 billion in damages), natural disasters typically affect GDP through infrastructure damage and reduced productivity. The earthquake (GDP growth rate from 5.2% to 2.8%) caused substantial economic disruption to affected regions.

P-RAG: The February 2023 earthquakes in Turkey caused severe economic damage worth billions of dollars. The disaster impacted various sectors including construction, manufacturing, and services. GDP growth was negatively affected from projected levels to around 2.8%, though this represents only one assessment among different economic evaluations.

GuaranTRAG (Ours): The February 2023 earthquakes in Turkey significantly impacted the country’s economic performance. According to the World Bank’s post-disaster assessment, the earthquakes caused \$84.1 billion in damages and reduced Turkey’s GDP growth rate from the initially projected 5.2% to 2.8% for 2023. This represents a substantial economic setback, with the disaster affecting infrastructure, productivity, and requiring extensive reconstruction efforts across the affected regions.

Table 7: Case study comparing different RAG approaches on a knowledge-intensive query. Red text shows how baseline methods fail to integrate factual information. Blue text represents coherent reasoning from parametric knowledge, while green text shows successfully integrated external factual information in our GUARANTRAG approach.

4.4 Case Study

Table 7 shows the case on different RAG method on Qwen3-14B without thinking mode. The existing methods cannot fully incorporate the retrieved documents while ours insert the necessary information into the appropriate position. More case study examples can be found in Appendix D.

5 Conclusion

In this paper, we addressed the critical knowledge integration problem in RAG systems, where conflicts between parametric and non-parametric knowledge sources significantly limit RAG effectiveness for knowledge-intensive applications. We introduced GUARANTRAG, a novel framework that resolves these conflicts through generating complementary inner-answers and refer-answers, then optimally fusing them using our proposed DPO-based training strategy and segment-level fusion mechanism. Comprehensive evaluations across five knowledge-intensive QA benchmarks

demonstrated that our approach consistently outperforms both conventional and dynamic RAG baselines, improving answer accuracy by up to 12.1% while reducing hallucination by 16.3%, with particularly strong performance on complex queries.

Limitations

While GUARANTRAG demonstrates significant improvements over existing RAG approaches, several limitations merit consideration. The additional computational cost introduced by our dual-answer generation and joint decoding mechanism may increase inference latency compared to standard RAG systems, potentially impacting real-time applications. Furthermore, the effectiveness of our approach may vary across different knowledge domains, particularly those with highly specialized terminology or complex relational structures not well-represented in our training data. The DPO-based training strategy, while effective, requires careful hyperparameter tuning to balance factual accuracy against answer conciseness. Our current implementation also assumes the availability of high-quality retrieval results, which may not always be guaranteed in real-world scenarios with ambiguous queries or limited knowledge sources. Despite these limitations, our experimental results confirm that GUARANTRAG’s benefits in factual accuracy and hallucination reduction substantially outweigh these modest constraints. Besides, our token-level fusion requires hidden-state access unavailable in closed-source API models; we approximate it by converting the soft intervention into temporary prompt-injected instructions, sampling multiple candidate fragments, and using a local lightweight model to score and select the best alignment. This approximation preserves segment-level evidence grounding but incurs performance gap and inference time costs versus full hidden-state fusion.

Ethical Considerations

Despite improving factuality, our system may still propagate biases or inaccuracies present in retrieved documents, particularly as the seamless integration of knowledge may obscure information provenance. The enhanced fluency and factual coherence could potentially make AI-generated misinformation more convincing if misused, necessitating appropriate safeguards in deployed applications.

Acknowledgements

We thank the reviewers, the AC, and the SAC for their constructive comments. This work is partially supported by Hong Kong RGC GRF No.14206324, Beijing Natural Science Foundation (Grant number: 4262075), Research Funds for NSD Construction, University of International Relations (Grant numbers: 3262026T23).

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Self-reflective retrieval augmented generation. In *NeurIPS 2023 workshop on instruction tuning and instruction following*.
- Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.
- Shiting Chen, Zijian Zhao, and Jinsong Chen. 2025. Each to their own: Exploring the optimal embedding in rag. *arXiv preprint arXiv:2507.17442*.
- Yuyan Chen, Yanghua Xiao, and Bang Liu. 2022. Grow-and-clip: Informative-yet-concise evidence distillation for answer explanation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 741–754. IEEE.
- Yang Deng, Wenxuan Zhang, Yaliang Li, Min Yang, Wai Lam, and Ying Shen. 2020. Bridging hierarchical and sequential context modeling for question-driven extractive answer summarization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1693–1696.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.
- Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. 2025. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 4206–4225.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.
- Yixiong Fang, Tianran Sun, Yuling Shi, and Xiaodong Gu. 2025. Attentionrag: Attention-guided context pruning in retrieval-augmented generation. *arXiv preprint arXiv:2503.10720*.
- Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu. 2025. Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *arXiv preprint arXiv:2504.14891*.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.
- Kazuki Hayashi, Hidetaka Kamigaito, Shinya Kouda, and Taro Watanabe. 2025. Iterkey: Iterative keyword generation with llms for enhanced retrieval augmented generation. *arXiv preprint arXiv:2505.08450*.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7029–7043.
- Wenqi Jiang, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska. 2024. Piperag: Fast retrieval-augmented generation via algorithm-system co-design. *arXiv preprint arXiv:2403.05676*.
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023a. **Active retrieval augmented generation**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.
- Rishi Kalra, Zekun Wu, Ayesha Gulley, Airlie Hilliard, Xin Guan, Adriano Koshiyama, and Philip Colin Treleaven. 2024. Hypa-rag: A hybrid parameter adaptive retrieval-augmented generation system for ai legal and policy applications. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 237–256.
- Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2024. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 23–35.

- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Alex Laitenberger, Christopher D. Manning, and Nelson F. Liu. 2025. Stronger baselines for retrieval-augmented generation with long-context language models. *arXiv preprint arXiv:2506.03989*.
- Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. Splade-v3: New baselines for splade. *arXiv preprint arXiv:2403.06789*.
- Ruizhe Li, Chen Chen, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. 2025. Attributing response to context: A jensen-shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. *arXiv preprint arXiv:2505.16415*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.
- Xin Lin, Zhenya Huang, Zhiqiang Zhang, Jun Zhou, and Enhong Chen. 2025. Explore what llm does not know in complex question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24585–24594.
- Craig Macdonald, Jinyuan Fang, Andrew Parry, and Zaiqiao Meng. 2025. Constructing and evaluating declarative rag pipelines in pyterrier. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4035–4040.
- Tommy Mordo, Moshe Tennenholtz, and Oren Kurland. 2024. Sponsored question answering. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 167–173.
- Zackary Rackauckas. 2024. Rag-fusion: a new take on retrieval-augmented generation. *International Journal on Natural Language Computing (IJNLC)*, 13.
- Zackary Rackauckas, Arthur Câmara, and Jakub Zavrel. 2024. Evaluating rag-fusion with ragelo: an automated elo-based framework. In *LLM4Eval@ SIGIR*.
- Mohammad Reza Rezaei and Adji Bousso Dieng. 2025. Vendi-rag: Adaptively trading-off diversity and quality significantly improves retrieval augmented generation with llms. *arXiv preprint arXiv:2502.11228*.
- Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.
- Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.
- Shangeetha Sivasothy, Scott Barnett, Stefanus Kurniawan, Zafaryab Rasool, and Rajesh Vasa. 2024. Rag-probe: An automated approach for evaluating rag applications. *arXiv preprint arXiv:2409.19019*.
- Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013.
- Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. 2025. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1240–1250.
- Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.
- Yujie Wang, Hu Zhang, Jiye Liang, and Ru Li. 2023. Dynamic heterogeneous-graph reasoning with language models and knowledge representation learning for commonsense question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14048–14063.
- Zilong Wang, Zifeng Wang, Long Le, Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, et al. 2024. Speculative rag: Enhancing retrieval augmented generation through drafting. In *The Thirteenth International Conference on Learning Representations*.
- Di Wu, Wasi Uddin Ahmad, Dejiao Zhang, Murali Krishna Ramanathan, and Xiaofei Ma. 2024. Repoformer: selective retrieval for repository-level code completion. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53270–53290.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, et al. 2024. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, et al. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*.

A Preliminary Experiments

A.1 Experimental Setup

To investigate the knowledge integration challenges in RAG systems, we conducted a systematic preliminary study using 500 knowledge-intensive queries drawn from a combination of established datasets: NaturalQuestions (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018). We specifically selected queries requiring factual information that might challenge the parametric knowledge of state-of-the-art LLMs. For our experiments, we used GPT-4o as the base language model and implemented a standard RAG pipeline with BM25 for retrieval. For each query, we retrieved the top-5 relevant documents from Wikipedia and varied the influence of these documents on the final response by systematically adjusting the retrieval weight parameter $\lambda \in [0.2, 0.4, 0.6, 0.8, 1.0]$. This parameter controls the balance between the model’s parametric knowledge and the information from retrieved documents during response generation, following the formulation:

$$P(y|q) = (1 - \lambda) \cdot P_{\theta}(y|q) + \lambda \cdot P_{\phi}(y|q, D) \quad (7)$$

where q is the query, y is the generated response, D represents the retrieved documents, $P_{\theta}(y|q)$ is the

probability from the base LLM, and $P_{\phi}(y|q, d)$ is the probability influenced by retrieved documents.

A.2 Evaluation Methodology

We employed two primary evaluation dimensions to measure the trade-off between factual correctness and response coherence:

Factual Accuracy. We evaluated factual correctness through a combination of automated and human evaluation:

- **Automated Factual Verification:** We used a fact-checking model trained on FEVER (Thorne et al., 2018) to verify claims in the generated responses against the retrieved documents, producing a factual correctness score.
- **Human Evaluation:** We recruit three PhD students independently assessed the factual accuracy of 100 randomly sampled responses on a 5-point scale, with 5 representing completely accurate responses. The inter-annotator agreement measured by Fleiss’ kappa was 0.78, indicating substantial agreement.

Coherence Quality. To evaluate the structural coherence and readability of responses, we measured:

- **Discourse Coherence:** Using GPT-4 to assess the logical flow and connection between sentences. The score rated on 5-point scale.
- **Linguistic Quality:** Human evaluators rated responses on a 1-5 scale for grammaticality, non-redundancy, referential clarity, and structural organization.
- **Self-consistency:** We checked for internal contradictions within responses, which often emerge when information from retrieved documents conflicts with the model’s parametric knowledge.

Additionally, we manually analyzed responses to identify “critical integration failures”, defined as instances where the RAG system either: (1) Failed to incorporate key factual information from retrieved documents; (2) Introduced contradictions between retrieved information and generated content; (3) Blended facts from multiple documents incorrectly; or (4) Prioritized less relevant information while omitting critical facts.

A.3 Results and Analysis

Our experimental results, summarized in Table 8. These preliminary findings highlight a critical limitation in current RAG architectures: they lack mechanisms to effectively reconcile parametric and

Retrieval Weight	Factual Accuracy	Coherence Quality	Integration Failures
0.2	67.1%	92.3%	71.4%
0.4	73.2%	84.7%	69.2%
0.6	76.8%	78.5%	67.8%
0.8	79.4%	72.8%	66.5%
1.0	81.5%	66.3%	67.3%

Table 8: Impact of retrieval weight on factual accuracy, coherence quality, and integration failures across 500 knowledge-intensive queries.

Failure Type	Percentage (%)
Information Omission	42.8
Contradictions	24.3
Fact Blending	19.1
Relevance Errors	13.8

Table 9: Distribution of knowledge integration failure types across the 67.3% of problematic RAG responses. The percentages represent the proportion of each failure type within the set of integration failures.

non-parametric knowledge sources. The observed trade-off between factual accuracy and coherence presents a significant challenge for real-world applications, where both qualities are essential for user trust and information utility.

A.4 Analysis of Knowledge Integration Failures

To gain deeper insights into the nature of knowledge integration failures, we manually categorized each failure instance from our analysis of the 67.3% problematic responses, shown in Table 9.

B Experimental Details

B.1 Datasets

We conducted experiments on five widely-used question answering benchmarks with varying reference document characteristics. Natural Questions (NQ) (Kwiatkowski et al., 2019) contains factoid questions derived from Google search queries with short reference passages extracted from Wikipedia articles, making it suitable for evaluating straightforward factual retrieval and integration. TruthfulQA (Lin et al., 2022) was specifically designed to evaluate factual accuracy and a model’s ability to avoid generating misinformation, containing questions across 38 categories where models might have a tendency to produce false answers, with

short-to-medium reference passages providing accurate information.

Wizard of Wikipedia (WoW) (Dinan et al.) features conversational questions requiring broader knowledge integration, including medium-length Wikipedia passages that models must effectively incorporate into coherent responses. HotpotQA (Yang et al., 2018) focuses on multi-hop reasoning questions that require synthesizing information across multiple documents, with reference passages of medium length that test a model’s ability to connect facts from different sources. ELI5 (Fan et al., 2019) consists of complex "explain like I’m five" questions from Reddit that require long-form explanations, with extensive reference documents that demand sophisticated knowledge integration to produce comprehensive yet accessible answers. This diverse selection enables us to evaluate our framework’s effectiveness across different knowledge integration scenarios, from straightforward factoid retrieval to complex multi-document reasoning tasks.

B.2 Retrieval Methods

We implemented four distinct retrieval approaches to ensure comprehensive evaluation. BM25 (Robertson and Walker, 1994) is a classic sparse retrieval method that relies on lexical matching between query terms and documents, using term frequency and inverse document frequency statistics to rank documents. RetroMAE (Xiao et al., 2022) is a masked autoencoder-based dense retriever that captures semantic relationships between queries and documents through contextual embeddings, enabling matching beyond exact lexical overlap.

SPLADE-v3 (Lassance et al., 2024) represents a state-of-the-art sparse-dense hybrid approach that combines lexical and semantic matching through learned sparse expansions of queries and documents, offering robust retrieval performance across diverse query types. HyDE (Gao et al., 2023) implements a generative retrieval approach that creates hypothetical document representations from queries before retrieval, effectively bridging the gap between query and document spaces for improved semantic matching. For fair comparison, all retrievers are configured to return the top-5 most relevant documents, which are then processed through our contrastive fusion mechanism with a relevance threshold of $\gamma = 0.68$ determined through validation experiments.

B.3 Evaluation Metrics

We evaluate our approach using metrics across three distinct levels of assessment. At the lexicon level, we employ Match, which evaluates the correctness of extracted answer spans against reference answers, focusing on exact matches for factual precision. ROUGE-L measures lexical overlap by focusing on the longest common subsequence between generated and reference texts, capturing fluency while allowing for word order variations. BLEU-4 assesses n-gram precision with emphasis on longer sequences to evaluate fluency and adequacy of generated responses relative to references.

At the semantic level, we utilize BERTScore, which leverages contextual embeddings to capture semantic similarity beyond surface forms, computing token similarities between candidate and reference texts using BERT representations. BEM (BERT-based Exact Match) applies BERT representations to determine semantic equivalence between responses and references, capturing meaning preservation even with lexical variation.

To evaluate the reference-document usage, we propose the following metrics.

(a) Hallucination Rate (Hal.) The hallucination rate quantifies the percentage of generated statements that contain factually incorrect information or assertions unsupported by the retrieved documents. To compute this metric, we extract atomic claims from model outputs and classify each as either factually supported or hallucinated:

$$\text{Hal.} = \frac{\text{Number of hallucinated claims}}{\text{Total number of claims}} \times 100\% \quad (8)$$

Each generated response is decomposed into atomic claims using a rule-based parser. Three expert annotators then independently verify each claim against the reference documents, with a claim marked as hallucinated if it either contradicts the reference information or makes assertions not present in the references. The final hallucination rate is the average across all test samples, with lower values indicating better performance.

(b) Entity Precision (Ent.) Entity precision measures how accurately a model reproduces entities from reference documents in its response. We extract named entities from both the reference documents and generated responses using a BERT-based named entity recognition system. The metric is

computed as:

$$\text{Ent.} = \frac{|\mathcal{E}_{\text{gen}} \cap \mathcal{E}_{\text{ref}}|}{|\mathcal{E}_{\text{gen}}|} \times 100\% \quad (9)$$

where \mathcal{E}_{gen} represents the set of entities in the generated response and \mathcal{E}_{ref} represents the set of entities in the reference documents. Entity matching accounts for variations in surface forms using a combination of exact matching, lemmatization, and semantic similarity. Higher values indicate better performance, reflecting more accurate entity reproduction.

(c) Structure Coherence (Struc.) Structure coherence evaluates the logical organization and readability of generated responses on a 5-point Likert scale. Unlike the other metrics, this is partially subjective and assessed through human evaluation. The scale is defined as follows:

$$\text{Struc.} = \frac{1}{N} \sum_{i=1}^N s_i \quad (10)$$

where N is the number of evaluated responses and s_i is the coherence score for response i , determined according to the following criteria: 1) Incoherent, disorganized response with no logical structure. 2) Poor organization with frequent logical breaks. 3) Acceptable structure with occasional inconsistencies. 4) Well-structured response with clear organization. 5) Excellent structure with logical flow and coherent paragraphs.

Each response is rated independently by three evaluators following a detailed rubric, and scores are averaged. This metric captures how well the model maintains logical structure while incorporating factual information from references, with higher values indicating better performance.

B.4 Baselines

We conduct comprehensive experiments to evaluate GUANTRAG against state-of-the-art baselines across multiple model scales. For backbone models, we employ Qwen-4B, Qwen-7B, and Qwen-14B as the foundation models for all experiments to enable fair comparisons across different parameter scales. We establish performance without retrieval augmentation for each model size as our base comparison.

For RAG-enhanced baselines, we implement Standard RAG, the conventional approach that directly incorporates retrieved passages into the

prompt before generation. SelfRAG (Asai et al., 2023) employs self-reflection mechanisms to critique and refine generated responses based on retrieved information, iteratively improving answer quality. RQ-RAG (Chan et al., 2024) utilizes retrieval quality estimation to dynamically adjust the influence of retrieved information based on confidence scores, balancing between parametric and retrieved knowledge. SOLAR (Kim et al., 2024) leverages strategic optimization of retrieval-augmented language models with a focus on improving the integration of external knowledge during generation.

We implement our GUARANTRAG framework with each of the Qwen models (4B, 7B, and 14B) to enable fair comparison across different parameter scales. This experimental design allows us to analyze both the effect of model scale and the effectiveness of various RAG techniques within a controlled environment.

B.5 Implementation Details

The knowledge decision module is distilled into a Qwen3-1.7B parameter model for efficiency. This module is trained on a custom dataset of labeled queries to predict whether external retrieval is necessary based on query characteristics. For contrastive fusion, we employ a 6-layer transformer encoder initialized with the weights from the upper layers of the base LLM. This component is optimized using AdamW with a learning rate of 2×10^{-5} and weight decay of 0.01. The training process involves 3 epochs on a curriculum of $\sim 50K$ synthetic query-answer pairs, with batch size 32.

We set the hyperparameters $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$ for the initial training phase, gradually shifting to $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ as training progresses. This dynamic adjustment helps balance factual accuracy against structural coherence as the model learns to better integrate knowledge. For external knowledge sources, we use a combination of Wikipedia (January 2023 snapshot), Wikidata, and ArXiv abstracts. These sources are indexed using Faiss to enable efficient retrieval with approximately 21M documents in the index.

The segment decomposition employs a discourse parsing technique with a threshold-based approach that identifies natural semantic boundaries in text. This results in an average of 5.7 segments per answer with average segment length of 43.2 tokens. The segmentation allows for more fine-grained fusion of information between the inner-answer and

refer-answer.

C Further Experiments

C.1 Detailed Experimental Results

This appendix provides comprehensive experimental results across all datasets, models, retrievers, and evaluation metrics.

C.1.1 Per-Dataset Performance Analysis

Tables 10 to Table 14 present detailed results for each dataset across all evaluation metrics. Our evaluation employs five complementary metrics: Match, ROUGE-L, BLEU-4, BERTScore, and BEM (BERT-based Evaluation Metric), providing a comprehensive assessment of both lexical and semantic similarity.

The results reveal several important patterns across datasets. First, GUARANTRAG demonstrates consistent improvements across all evaluation metrics, with the largest gains observed in exact match scores, indicating superior factual accuracy. Second, performance improvements are particularly pronounced on complex reasoning tasks like HotpotQA, where our approach achieves up to 8.7 percentage points improvement in exact match over P-RAG. Third, the benefits of our approach scale positively with model capacity, suggesting that stronger parametric knowledge enables more effective knowledge integration.

C.1.2 API-Based Model Evaluation

To assess the generalizability of our approach beyond self-hosted models, we evaluate GUARANTRAG on state-of-the-art API-based language models. Table 15 presents comprehensive results across four leading commercial models.

The results demonstrate that GUARANTRAG consistently improves performance across all API-based models, achieving competitive results with state-of-the-art systems.

C.1.3 Retriever Comparison Analysis

Table 16 presents comprehensive results across different retrieval methods for Qwen3-14B-Thinking. The consistency of improvements across diverse retrieval mechanisms validates that our approach addresses fundamental knowledge integration challenges rather than retrieval-specific issues.

Dense retrievers generally outperform sparse methods, with RetroMAE showing the strongest performance. However, GUARANTRAG maintains

Method	Match	ROUGE-L	BLEU-4	BERTScore	BEM
Qwen3-8B	34.2	68.7	42.1	78.4	76.5
w/ Standard RAG	38.6	72.3	46.8	81.2	78.1
w/ SelfRAG	42.1	75.9	51.2	83.7	80.6
w/ RQ-RAG	43.8	77.1	52.9	84.6	81.1
w/ SOLAR	45.3	78.4	54.7	85.2	82.4
w/ DA-RAG	40.7	74.8	49.1	82.9	79.5
w/ FLARE	44.2	77.8	53.4	84.7	81.9
w/ DRAGIN	44.6	78.2	53.8	85.1	82.3
w/ P-RAG	47.1	80.6	56.9	86.4	83.0
w/ GuarantRAG	51.8	83.2	61.4	88.7	86.9
Qwen3-14B	36.8	70.9	44.7	79.6	78.0
w/ Standard RAG	41.4	74.8	49.3	82.7	80.8
w/ SelfRAG	45.2	78.6	54.1	85.1	82.4
w/ RQ-RAG	47.1	80.2	55.8	86.0	83.4
w/ SOLAR	48.9	81.7	58.2	86.9	84.3
w/ DA-RAG	43.6	77.4	51.9	84.2	81.9
w/ FLARE	47.8	80.9	56.7	86.3	83.8
w/ DRAGIN	48.2	81.3	57.1	86.7	84.1
w/ P-RAG	50.4	83.6	60.2	87.8	85.0
w/ GuarantRAG	54.9	86.1	64.7	90.3	89.5
Qwen3-14B-T	38.1	72.4	46.2	80.8	79.5
w/ Standard RAG	43.7	76.3	51.1	83.9	82.1
w/ SelfRAG	47.8	80.1	56.4	86.4	84.3
w/ RQ-RAG	49.6	81.7	58.2	87.2	85.5
w/ SOLAR	51.3	83.2	60.9	88.1	86.0
w/ DA-RAG	46.2	79.1	54.7	85.3	83.7
w/ FLARE	50.1	82.4	59.1	87.6	85.6
w/ DRAGIN	50.5	82.8	59.5	88.0	86.1
w/ P-RAG	52.8	85.1	62.7	89.2	87.4
w/ GuarantRAG	56.7	87.9	67.1	91.8	91.0

Table 10: Detailed results on Natural Questions dataset with BM25 retriever. All scores are percentages except BERTScore which is on 0-1 scale ($\times 100$).

consistent improvements across all retrieval methods. Notably, our approach also reduces hallucination rates across all retrievers, with the most significant reduction observed with RetroMAE.

C.1.4 Statistical Significance Analysis

To ensure the reliability of our results, we conduct statistical significance testing using paired bootstrap resampling with $n = 10$ iterations. Table 17 presents p-values for key comparisons.

All improvements achieved by GUARANTRAG are statistically significant at the $p < 0.002$ level, providing strong evidence for the effectiveness of our approach.

C.2 Settings on Input Length Analysis

To provide a comprehensive evaluation of GUARANTRAG’s robustness across varying input conditions, we systematically categorize queries and reference documents along multiple dimensions of complexity. This analysis enables us to understand how our framework performs under different knowledge integration scenarios and computational

demands.

Query Length Categorization. We partition queries into three categories based on word count: *short queries* contain fewer than 10 words and typically represent straightforward factual questions (e.g., “What is the capital of France?”); *medium queries* range from 10-25 words and often involve more specific information requests with contextual constraints (e.g., “What are the environmental impacts of solar panel manufacturing and disposal?”); and *long queries* exceed 25 words, usually comprising complex, multi-part questions requiring detailed explanations (e.g., “Explain the differences between classical and quantum computing architectures, including their respective advantages for cryptographic applications”).

Reasoning Complexity Classification. We classify queries into three reasoning complexity levels based on the cognitive processes required for answering: *Simple reasoning* queries require direct fact retrieval or single-step inference from the knowledge base; *moderate reasoning* queries in-

Method	Match	ROUGE-L	BLEU-4	BERTScore	BEM
Qwen3-8B	28.4	63.2	38.7	74.1	73.8
w/ Standard RAG	32.1	67.4	42.9	77.3	75.8
w/ SelfRAG	35.8	71.2	47.1	80.6	78.3
w/ RQ-RAG	37.2	72.6	48.4	81.4	79.2
w/ SOLAR	38.9	74.1	50.2	82.7	80.1
w/ DA-RAG	34.6	70.1	45.8	79.8	77.7
w/ FLARE	37.6	73.4	49.1	81.9	79.5
w/ DRAGIN	38.1	73.8	49.6	82.3	79.9
w/ P-RAG	40.4	75.9	52.7	83.6	81.4
w/ GuarantRAG	44.2	78.7	57.3	86.1	84.6
Qwen3-14B	30.7	65.8	41.2	76.4	75.4
w/ Standard RAG	34.9	70.1	45.8	79.7	78.2
w/ SelfRAG	38.7	74.2	50.4	82.9	80.8
w/ RQ-RAG	40.3	75.8	52.1	83.8	81.6
w/ SOLAR	42.1	77.4	54.3	84.9	82.8
w/ DA-RAG	37.2	73.1	48.7	81.7	80.1
w/ FLARE	40.9	76.7	53.2	84.2	82.1
w/ DRAGIN	41.4	77.1	53.7	84.6	82.5
w/ P-RAG	43.8	79.3	56.4	85.8	83.6
w/ GuarantRAG	47.6	82.1	60.9	88.3	87.1
Qwen3-14B-T	32.4	67.9	43.1	78.2	76.8
w/ Standard RAG	36.2	72.1	47.6	81.4	79.8
w/ SelfRAG	40.6	76.4	52.8	84.7	82.7
w/ RQ-RAG	42.3	78.1	54.6	85.6	83.8
w/ SOLAR	44.2	79.7	57.1	86.7	84.9
w/ DA-RAG	39.1	75.2	51.4	83.6	81.9
w/ FLARE	43.0	78.9	55.8	86.1	84.2
w/ DRAGIN	43.5	79.3	56.3	86.5	84.6
w/ P-RAG	45.7	81.6	59.4	87.8	85.9
w/ GuarantRAG	49.8	84.3	63.2	90.1	88.9

Table 11: Detailed results on TruthfulQA dataset with BM25 retriever.

involve multi-step logical connections or require synthesizing information from multiple sources; and *complex reasoning* queries demand sophisticated analytical thinking, causal reasoning, or multi-hop inference across diverse knowledge domains. This classification is performed using a combination of automated linguistic features (syntactic complexity, semantic diversity) and manual verification on a subset of queries.

Reference Document Length Segmentation.

Retrieved documents are categorized by token count to assess knowledge integration performance across varying information density: *concise documents* contain fewer than 300 tokens and typically provide focused, specific information; *moderate documents* range from 300-800 tokens and offer comprehensive coverage of topics with multiple supporting details; and *extensive documents* exceed 800 tokens, containing rich contextual information that requires sophisticated filtering and integration strategies. Token counting is performed using the Qwen3 tokenizer to ensure consistency with the underlying language model.

C.3 Analysis on Different Answer Fusion Mechanisms

We conduct a comprehensive analysis of different mechanisms for fusing the inner-answer and refer-answer in our GUARANTRAG framework. Our proposed joint decoding mechanism dynamically integrates knowledge during token-by-token generation, but alternative fusion strategies deserve investigation to validate our design choices.

Fusion Strategy Variants. We compare four distinct fusion mechanisms: (1) **Prompt-based Fusion**: concatenating inner-answer and refer-answer in a prompt template and instructing the LLM to generate a fused response; (2) **Mean Segment Representation**: replacing our final-token segment representation with mean-pooled token embeddings within each segment; (3) **Attention-based Fusion**: using cross-attention mechanisms to weight and combine representations from both answers; and (4) our proposed **Joint Decoding** approach.

Prompt-based Fusion Analysis. This approach uses an early fusion mechanism that integrates documents at an earlier stage, where we provide both

Method	Match	ROUGE-L	BLEU-4	BERTScore	BEM
Qwen3-8B	31.7	66.4	40.3	75.8	74.8
w/ Standard RAG	35.8	70.7	44.9	78.9	77.2
w/ SelfRAG	39.6	74.8	49.7	82.1	79.8
w/ RQ-RAG	41.2	76.3	51.4	83.0	80.6
w/ SOLAR	42.7	77.9	53.2	84.1	81.9
w/ DA-RAG	38.1	73.6	47.8	81.2	78.7
w/ FLARE	41.9	76.8	52.1	83.4	80.8
w/ DRAGIN	42.4	77.2	52.6	83.8	81.2
w/ P-RAG	44.8	79.7	55.9	85.2	82.4
w/ GuarantRAG	48.9	82.6	60.4	87.3	85.3
Qwen3-14B	33.9	68.7	42.8	77.3	76.0
w/ Standard RAG	38.2	73.1	47.6	80.6	78.5
w/ SelfRAG	42.1	77.4	52.5	83.7	81.3
w/ RQ-RAG	43.8	79.0	54.3	84.6	82.1
w/ SOLAR	45.6	80.7	56.8	85.8	83.5
w/ DA-RAG	40.7	76.2	50.4	82.8	80.1
w/ FLARE	44.3	79.6	55.1	84.9	82.6
w/ DRAGIN	44.8	80.0	55.6	85.3	83.0
w/ P-RAG	47.3	82.4	58.9	86.7	84.2
w/ GuarantRAG	51.6	85.3	63.7	89.1	87.3
Qwen3-14B-T	34.8	69.4	43.7	78.1	76.5
w/ Standard RAG	39.1	73.8	48.2	81.3	79.1
w/ SelfRAG	43.2	78.1	53.4	84.6	82.2
w/ RQ-RAG	45.0	79.7	55.2	85.5	83.0
w/ SOLAR	46.9	81.4	57.8	86.7	84.3
w/ DA-RAG	41.8	77.0	51.9	83.7	81.4
w/ FLARE	45.7	80.3	56.1	85.8	83.5
w/ DRAGIN	46.2	80.7	56.6	86.2	83.9
w/ P-RAG	48.6	83.1	59.8	87.6	85.1
w/ GuarantRAG	52.8	86.2	64.9	90.4	88.7

Table 12: Detailed results on Wizard of Wikipedia dataset with BM25 retriever.

answers to the LLM with the instruction: “Given the following parametric answer and reference-based answer, generate an optimal response that combines the best aspects of both.” While conceptually straightforward, this method suffers from several limitations: (1) increased computational overhead due to longer input sequences, (2) potential confusion when the two answers contain contradictory information, and (3) limited control over the fusion granularity. Our experiments show this approach achieves 67.6% average performance, underperforming by 4.1% compared to our joint decoding mechanism.

Segment Representation Analysis. Mean aggregation merges subword tokens into a single representation by averaging their vector embeddings. We implement this by computing $h(s) = \frac{1}{|s|} \sum_{i=1}^{|s|} h_i$ for each segment s , where h_i represents individual token embeddings. While this provides a more comprehensive segment representation, it introduces noise from less informative tokens and dilutes the significance of semantically crucial final tokens. The mean representation

approach achieves 69.7% average performance, demonstrating that final-token representation better captures segment-level semantics for our fusion task.

Attention-based Fusion Analysis. We implement a cross-attention mechanism that computes attention weights between inner-answer and refer-answer segments: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where queries come from the current generation context and keys and values from refer-answer segments. While this approach shows improved performance (71.0% average), it requires additional parameters and computational overhead during inference. Moreover, the attention mechanism struggles to maintain temporal coherence during sequential generation, leading to inconsistent fusion decisions across decoding steps.

Joint Decoding Advantages. Our joint decoding mechanism outperforms alternative approaches by: (1) preserving the natural generation flow of the base model while selectively incorporating external knowledge, (2) using computationally efficient cosine similarity for segment matching, (3)

Method	Match	ROUGE-L	BLEU-4	BERTScore	BEM
Qwen3-8B	24.6	59.8	35.2	71.4	72.5
w/ Standard RAG	28.2	64.1	39.7	74.8	75.7
w/ SelfRAG	31.9	68.4	44.3	78.2	78.1
w/ RQ-RAG	33.6	70.1	46.1	79.1	79.0
w/ SOLAR	35.2	71.7	48.0	80.5	80.1
w/ DA-RAG	30.4	67.2	42.8	77.3	77.2
w/ FLARE	34.1	70.8	47.2	79.7	79.4
w/ DRAGIN	34.6	71.2	47.7	80.1	79.8
w/ P-RAG	36.9	73.6	50.4	81.8	81.1
w/ GuarantRAG	40.7	77.1	55.2	84.6	84.9
Qwen3-14B	26.4	62.1	37.6	73.8	74.6
w/ Standard RAG	30.6	66.7	42.3	77.1	77.3
w/ SelfRAG	34.5	71.2	47.1	80.4	79.8
w/ RQ-RAG	36.3	72.9	49.0	81.4	80.8
w/ SOLAR	38.1	74.6	51.2	82.7	81.9
w/ DA-RAG	32.8	69.5	45.4	79.2	78.7
w/ FLARE	36.9	73.6	50.1	81.9	80.6
w/ DRAGIN	37.4	74.0	50.6	82.3	81.0
w/ P-RAG	39.8	76.4	53.7	84.1	82.5
w/ GuarantRAG	43.9	79.8	58.4	86.9	86.0
Qwen3-14B-T	30.2	67.9	42.8	78.6	79.0
w/ Standard RAG	34.1	72.4	47.6	81.9	81.7
w/ SelfRAG	38.4	77.1	52.9	85.2	84.3
w/ RQ-RAG	40.3	79.0	55.0	86.3	85.4
w/ SOLAR	42.2	80.8	57.4	87.6	86.5
w/ DA-RAG	36.8	75.4	51.1	83.9	83.1
w/ FLARE	41.1	79.7	56.2	86.8	85.7
w/ DRAGIN	41.6	80.1	56.7	87.2	86.1
w/ P-RAG	44.2	82.7	60.1	88.9	87.6
w/ GuarantRAG	48.9	86.2	64.8	91.4	90.7

Table 13: Detailed results on HotpotQA dataset with BM25 retriever. Note the particularly strong improvements on this multi-hop reasoning task.

maintaining consistent fusion decisions through explicit similarity computation, and (4) avoiding the need for additional trainable parameters during inference. The 2.7% average improvement over attention-based fusion validates our design choice for dynamic, similarity-driven knowledge integration. Besides, Joint decoding requires minimal additional computation, only similarity calculations, while prompt-based fusion increases input length by 2.3× on average, and attention-based fusion introduces 12% inference latency overhead. Our approach achieves superior performance with negligible computational cost, making it practical for deployment scenarios.

C.4 Analysis on Different Decision Making Module

While our framework achieves substantial improvements through joint decoding and dual-path answer generation, a critical question remains: how sensitive is GUARANTRAG to different decision-making mechanisms that determine whether to invoke retrieval? To demonstrate the decision-

making-agnostic nature of our approach, we conduct comprehensive experiments comparing various decision modules across different complexity levels.

Decision Module Variants. We evaluate five distinct decision-making strategies: (1) **Query-based Classifier:** Our proposed distilled LLM fine-tuned on 2,000 labeled queries. (2) **Keyword-based Filter:** A rule-based system using predefined temporal and entity keywords. (3) **Confidence-based Threshold:** Selecting queries where the base model’s generation confidence falls below a threshold of 0.7. (4) **Similarity-based Retrieval:** Using retrieval similarity scores with a threshold of 0.6 to determine knowledge necessity. (5) **Random Selection:** Randomly selecting 50% of queries for retrieval augmentation as a baseline control.

Table 19 reveals several key insights about decision module sensitivity. First, all decision variants achieve remarkably similar performance, with less than 1% variation despite significantly different query allocation strategies. This consistency

Method	Match	ROUGE-L	BLEU-4	BERTScore	BEM
Qwen3-8B	30.9	68.2	41.5	76.3	75.6
w/ Standard RAG	34.7	72.1	45.8	79.4	78.5
w/ SelfRAG	38.5	76.4	50.7	82.6	81.8
w/ RQ-RAG	40.1	77.9	52.4	83.5	82.6
w/ SOLAR	41.8	79.6	54.7	84.8	83.9
w/ DA-RAG	37.2	75.1	48.9	81.7	80.5
w/ FLARE	40.6	78.7	53.2	84.1	83.2
w/ DRAGIN	41.1	79.1	53.7	84.5	83.6
w/ P-RAG	43.4	81.8	56.9	86.2	85.2
w/ GuarantRAG	47.2	84.7	61.3	88.4	87.4
Qwen3-14B	32.6	70.5	43.9	78.1	77.9
w/ Standard RAG	36.8	74.6	48.2	81.3	80.4
w/ SelfRAG	40.9	78.9	53.4	84.4	83.6
w/ RQ-RAG	42.6	80.6	55.2	85.4	84.4
w/ SOLAR	44.4	82.3	57.8	86.7	85.8
w/ DA-RAG	39.3	77.6	51.7	83.5	82.3
w/ FLARE	43.2	81.4	56.1	85.9	84.9
w/ DRAGIN	43.7	81.8	56.6	86.3	85.3
w/ P-RAG	46.1	84.3	60.2	88.1	87.2
w/ GuarantRAG	49.8	87.1	64.7	90.2	89.2
Qwen3-14B-T	31.4	68.9	42.1	77.6	76.7
w/ Standard RAG	35.2	72.8	46.4	80.8	79.8
w/ SelfRAG	39.1	77.3	51.7	83.9	82.7
w/ RQ-RAG	40.8	79.0	53.5	84.9	83.6
w/ SOLAR	42.6	80.8	56.1	86.2	85.0
w/ DA-RAG	37.9	76.1	50.2	82.8	81.5
w/ FLARE	41.4	79.7	54.8	85.4	84.2
w/ DRAGIN	41.9	80.1	55.3	85.8	84.6
w/ P-RAG	44.1	82.6	58.4	87.5	86.1
w/ GuarantRAG	46.7	84.9	61.2	89.1	87.5

Table 14: Detailed results on ELI5 dataset with BM25 retriever. This dataset requires long-form explanatory responses.

Method	Performance Across Datasets					Average	Reference Usage		
	NQ	TruthfulQA	WoW	HotpotQA	ELI5		Hal.	Ent.	Struc.
DeepSeek-V3	77.2	72.1	75.8	69.4	76.3	74.2	16.8	–	4.81
w/ Standard RAG	79.8	74.6	78.3	72.1	79.1	76.8	14.2	85.3	4.47
w/ SelfRAG	81.4	76.2	80.1	73.8	80.9	78.5	12.9	87.1	4.54
w/ P-RAG	82.7	77.9	81.6	75.2	82.3	79.9	11.4	88.7	4.61
w/ GuarantRAG	84.1	79.3	83.2	76.8	83.7	81.4	9.6	90.4	4.68
DeepSeek-R1	78.9	73.8	77.1	71.2	77.6	75.7	15.3	–	4.84
w/ Standard RAG	81.2	76.1	79.4	73.7	80.0	78.1	13.1	86.2	4.49
w/ SelfRAG	82.6	77.8	81.0	75.3	81.7	79.7	11.8	87.9	4.57
w/ P-RAG	84.1	79.4	82.7	76.9	83.2	81.3	10.2	89.5	4.64
w/ GuarantRAG	85.7	81.1	84.3	78.4	84.8	82.9	8.4	91.3	4.71
Gemini-2.5-Pro	76.4	71.3	74.9	68.1	75.2	73.2	17.6	–	4.79
w/ Standard RAG	78.9	73.7	77.2	70.8	77.8	75.7	15.1	84.6	4.44
w/ SelfRAG	80.3	75.4	78.9	72.4	79.5	77.3	13.6	86.3	4.52
w/ P-RAG	81.8	77.1	80.5	74.1	81.0	78.9	12.0	87.9	4.59
w/ GuarantRAG	83.4	78.7	82.1	75.6	82.6	80.5	10.3	89.6	4.66
Claude-3.7-Sonnet	75.1	69.8	73.6	66.7	74.3	71.9	18.9	–	4.77
w/ Standard RAG	77.6	72.3	76.1	69.5	76.9	74.5	16.2	83.8	4.42
w/ SelfRAG	79.1	74.0	77.8	71.1	78.6	76.1	14.7	85.4	4.49
w/ P-RAG	80.7	75.8	79.4	72.8	80.1	77.8	13.1	87.1	4.56
w/ GuarantRAG	82.3	77.4	81.0	74.3	81.7	79.3	11.4	88.9	4.63

Hal.: Hallucination Rate (%); Ent.: Entity Precision (%); Struc.: Structure Coherence (5-point scale)

Table 15: Performance evaluation on API-based language models. All experiments use BM25 retriever with consistent evaluation protocols. Performance metrics represent averages across five evaluation measures.

Method	BM25		SPLADE-v3		RetroMAE		HyDE	
	Avg	Hal.	Avg	Hal.	Avg	Hal.	Avg	Hal.
Qwen3-14B-T	59.5	32.9	59.5	32.9	59.5	32.9	59.5	32.9
w/ Standard RAG	64.0	31.4	64.3	30.8	65.1	30.2	64.8	30.6
w/ SelfRAG	68.7	28.2	69.1	27.8	69.8	27.1	69.4	27.5
w/ RQ-RAG	70.4	26.9	70.8	26.4	71.6	25.8	71.2	26.1
w/ SOLAR	71.9	25.7	72.4	25.1	73.2	24.5	72.8	24.9
w/ DA-RAG	67.8	28.9	68.2	28.4	68.9	27.8	68.5	28.2
w/ FLARE	70.6	26.4	71.1	25.9	71.8	25.3	71.4	25.7
w/ DRAGIN	71.1	26.0	71.6	25.5	72.3	24.9	71.9	25.3
w/ P-RAG	72.8	24.6	73.3	24.1	74.1	23.5	73.7	23.9
w/ GuarantRAG	75.2	21.8	75.8	21.2	76.6	20.6	76.2	21.0

Table 16: Performance comparison across different retrievers using Qwen3-14B-T. “Avg” represents average performance across all datasets and metrics. “Hal.” represents hallucination rate (%).

Comparison	Average Score	p-value
GuarantRAG		
vs. P-RAG	+2.4%	< 0.002
vs. SOLAR	+3.3%	< 0.0015
vs. SelfRAG	+6.5%	< 0.002
vs. Standard RAG	+11.2%	< 0.001
vs. No RAG	+15.7%	< 0.001

Table 17: Statistical significance analysis comparing GuarantRAG against baselines across all datasets and metrics. All improvements are statistically significant at $\alpha = 0.001$.

demonstrates that GUARANTRAG’s effectiveness primarily stems from the joint decoding mechanism rather than precise decision-making accuracy. Second, different modules exhibit varying computational efficiency profiles, keyword-based filtering achieves the lowest latency and computational cost by avoiding neural inference, while confidence-based approaches incur higher overhead due to additional forward passes for uncertainty estimation.

C.5 Analysis on Thinking Mode of Qwen3

Comparing Qwen3-14B and Qwen3-14B-T results in Table 1, we observe that thinking mode consistently improves performance across all RAG methods, with the enhancement being most pronounced for complex reasoning tasks. Specifically, thinking mode provides an average improvement of 2.2% for standard methods, increasing to 3.8% for sophisticated approaches like P-RAG, and reaching 4.2% for our GUARANTRAG framework. This suggests that the deliberative reasoning process in thinking mode creates synergistic effects with advanced knowledge integration mechanisms.

Table 20 presents a detailed breakdown of think-

ing mode effects across different query complexities. For simple factoid queries, thinking mode shows modest improvements of 1.3-2.1%. However, for complex multi-hop reasoning tasks, the enhancement reaches 5.2-7.8%, indicating that deliberative reasoning is particularly beneficial when queries require sophisticated knowledge synthesis.

D Additional Case Studies

This appendix presents additional case studies comparing our GuarantRAG approach with baseline RAG methods. Table 21 to 25 highlight our method’s superior ability to integrate retrieved information while maintaining coherence and factual accuracy. The color-coding helps visualize how each approach handles knowledge integration: **red text** indicates failed integration in baselines, **blue text** shows coherent reasoning from parametric knowledge, and **green text** represents successfully integrated external factual information.

Fusion Method	NQ	TruthfulQA	HotpotQA	Avg.
Prompt-based Fusion	71.2	66.4	65.3	67.6
Mean Segment Repr.	73.8	68.1	67.2	69.7
Attention-based Fusion	74.9	69.3	68.7	71.0
Joint Decoding (Ours)	76.4	69.8	68.9	71.7

Table 18: Performance comparison of different answer fusion mechanisms on selected datasets using Qwen3-8B.

Decision Module	RAG %	Non-RAG %	Latency (ms)	Performance
Query-based Classifier	43.2	56.8	127.4	76.7
Keyword-based Filter	38.7	61.3	119.8	76.3
Confidence-based	52.1	47.9	139.6	76.4
Similarity-based	47.8	52.2	134.2	76.5
Random Selection	50.0	50.0	132.1	75.9

Table 19: Comparison of different decision-making modules on the combined test set. RAG % and Non-RAG % indicate the proportion of queries assigned to each processing path. Latency measures average response time per query. Performance represents the averaged score across all five datasets using our proposed joint decoding framework.

Method	Performance by Query Complexity				
	Simple (1-2)	Moderate (2-3)	Complex (3-4)	Multi-hop (4-5)	Reasoning (5)
<i>Qwen3-14B vs Qwen3-14B-T Performance Gap</i>					
Standard RAG	+1.3	+2.4	+3.8	+5.2	+6.1
SelfRAG	+1.7	+2.9	+4.2	+5.8	+6.7
RQ-RAG	+1.8	+3.1	+4.5	+6.0	+6.9
SOLAR	+1.9	+3.2	+4.7	+6.3	+7.2
P-RAG	+2.1	+3.5	+5.0	+6.8	+7.6
GuarantRAG	+2.4	+3.8	+5.4	+7.1	+7.8
<i>Thinking Mode Efficiency Analysis</i>					
Thinking Steps	2.3	4.7	8.2	12.4	16.8
Latency (s)	+0.8	+1.9	+4.1	+7.2	+11.3

Table 20: Performance improvement of thinking mode over standard mode across query complexities, with efficiency metrics. Values represent percentage improvements averaged across all five datasets.

Query: What effect does metformin have on cardiovascular outcomes in Type 2 diabetes patients?

Retrieved Document: A 2021 meta-analysis of 25 randomized controlled trials involving 126,478 patients found that metformin reduced major cardiovascular events by 19% (RR=0.81, 95% CI 0.73-0.90) in Type 2 diabetes patients compared to placebo. The UKPDS study specifically showed a 39% reduction in myocardial infarction risk over 10 years of follow-up. However, cardiovascular benefits were less pronounced in patients with established heart disease.

Standard RAG: Metformin has been shown to have beneficial effects on cardiovascular outcomes in Type 2 diabetes patients. Research (meta-analysis, 19% reduction) suggests it reduces cardiovascular events. It works through multiple mechanisms including improved insulin sensitivity (UKPDS study showed 39% MI reduction) and modest weight loss.

P-RAG: Metformin is considered the first-line medication for Type 2 diabetes and has been studied for its cardiovascular effects. Evidence suggests metformin provides cardiovascular protection beyond glycemic control. It appears to reduce the risk of cardiovascular events by approximately 19% in diabetic patients, though the exact mechanism remains under investigation.

GuarantRAG (Ours): Metformin, the first-line treatment for Type 2 diabetes, demonstrates significant cardiovascular protective effects. According to a 2021 meta-analysis of 25 randomized controlled trials with 126,478 patients, metformin reduced major cardiovascular events by 19% (RR=0.81) compared to placebo. The landmark UKPDS study specifically demonstrated a 39% reduction in myocardial infarction risk over 10 years of follow-up. These benefits likely stem from metformin's multiple mechanisms of action, including improved insulin sensitivity, modest weight reduction, and anti-inflammatory effects. However, it's worth noting that these cardiovascular benefits appear less pronounced in patients with established heart disease.

Table 21: Case study on metformin and cardiovascular outcomes in diabetes.

Query: How did the Meiji Restoration transform Japan's economy?

Retrieved Document: The Meiji Restoration (1868) radically transformed Japan's economy by implementing Western-style industrialization policies. The government established state-owned enterprises in strategic sectors, with shipbuilding yards increasing from 2 in 1868 to 38 by 1900. Japan's industrial production grew at 6.2% annually from 1868-1912, with textile exports rising from ¥0.5 million in 1868 to ¥236 million by 1900. The "shokusan kogyo" (promote industry) policy abolished feudal restrictions, while the 1872 Land Tax Reform created a modern taxation system, providing 80% of government revenue until 1890. Japan transitioned from a feudal agrarian economy to Asia's first industrialized nation within one generation.

Standard RAG: The Meiji Restoration of 1868 transformed Japan's economy from feudal to modern. The government implemented modernization policies (industrial production grew 6.2% annually) and promoted Western-style industrialization. This included building infrastructure (shipbuilding yards increased from 2 to 38) and reforming education. Japan rapidly industrialized as a result.

P-RAG: The Meiji Restoration of 1868 fundamentally transformed Japan's economy through deliberate modernization policies. The new government abolished feudal economic structures and implemented Western industrial models. Key policies included "shokusan kogyo" (promote industry) and the 1872 Land Tax Reform, which provided 80% of government revenue until 1890. These reforms enabled Japan's rapid industrialization and economic growth.

GuarantRAG (Ours): The Meiji Restoration of 1868 marked a pivotal turning point that fundamentally transformed Japan's economy. Under the "shokusan kogyo" (promote industry) policy, the government established state-owned enterprises in strategic sectors, exemplified by the increase in shipbuilding yards from just 2 in 1868 to 38 by 1900. Japan achieved remarkable industrial growth at 6.2% annually from 1868-1912, with textile exports skyrocketing from ¥0.5 million to ¥236 million by 1900. This rapid industrialization was facilitated by institutional reforms that dismantled feudal barriers to economic development. A critical component was the 1872 Land Tax Reform, which created a modern taxation system providing 80% of government revenue until 1890. Through these comprehensive measures, Japan accomplished what few nations had achieved—transitioning from a feudal agrarian economy to Asia's first industrialized nation within a single generation.

Table 22: Case study on the economic transformation during the Meiji Restoration.

Query: What mechanisms explain how quantum tunneling affects nuclear fusion in stars?

Retrieved Document: Nuclear fusion in stars occurs when hydrogen nuclei overcome their electrostatic repulsion to fuse into helium. Classical physics predicts this should be nearly impossible at stellar core temperatures (15 million K in the Sun) since particles need approximately 1 billion K to overcome the Coulomb barrier through thermal energy alone. Quantum tunneling allows particles to penetrate this barrier despite having insufficient energy, increasing fusion probability by a factor of 10^{20} . Specifically, the tunneling probability is proportional to $\exp(-2\pi Ze^2/hv)$, where Z represents nuclear charge, e is electron charge, h is Planck's constant, and v is relative velocity. This quantum effect enables the p-p chain reaction that powers main sequence stars, with tunneling enhancing the reaction rate enough to explain stellar lifespans of billions rather than millions of years.

Standard RAG: Quantum tunneling is a phenomenon that allows particles to pass through energy barriers they couldn't normally overcome. In stars, hydrogen nuclei need to overcome electrostatic repulsion to undergo fusion. Classical physics suggests temperatures of 1 billion K would be needed, but stellar cores are much cooler. Quantum tunneling enables fusion at these lower temperatures (increasing probability by a factor of 10^{20}).

P-RAG: Quantum tunneling plays a crucial role in stellar nucleosynthesis by enabling fusion reactions that would be nearly impossible under classical physics. This quantum mechanical effect allows particles to penetrate energy barriers despite having insufficient energy. In the Sun's core (15 million K), particles can quantum tunnel through the Coulomb barrier that would classically require 1 billion K to overcome. This tunneling effect explains why stars can sustain fusion reactions over billions of years.

GuarantRAG (Ours): Quantum tunneling fundamentally enables stellar fusion by providing a pathway for nuclear reactions that would be effectively impossible under classical physics. At the Sun's core temperature of 15 million K, hydrogen nuclei possess insufficient thermal energy to overcome the Coulomb barrier, which would classically require temperatures approaching 1 billion K. Quantum tunneling circumvents this limitation by allowing particles to penetrate this electrostatic barrier despite having inadequate energy, increasing fusion probability by a remarkable factor of 10^{20} . This quantum phenomenon operates through wave-particle duality, where nuclei behave as probability waves that can extend beyond classical boundaries. The tunneling probability follows the relationship $\exp(-2\pi Ze^2/hv)$, incorporating nuclear charge (Z), electron charge (e), Planck's constant (h), and relative velocity (v). This mechanism elegantly explains how main sequence stars can sustain fusion reactions for billions rather than millions of years, maintaining stable energy production through what appears to be an improbable quantum process.

Table 23: Case study on quantum tunneling in stellar fusion processes.

Query: How do environmental, social, and governance (ESG) factors impact corporate financial performance?

Retrieved Document: A 2021 meta-analysis by NYU Stern Center for Sustainable Business examined 1,141 peer-reviewed studies (2015-2020) on ESG-financial performance correlation. Results showed 58% of corporate studies found positive ESG-financial performance relationships, with only 8% negative. For investment studies, 59% showed positive performance from ESG portfolios vs. conventional, with just 14% negative. Environmental factors demonstrated the strongest positive correlation (environmental management systems linked to 12.9% higher ROA). Social practices showed a 7.5% boost to ROE, while governance delivered a 5.8% increase in Tobin's Q ratio. ESG impact varied by industry, with consumer goods and financial services seeing greater benefits (11.5% and 10.2% increased ROE respectively) than energy and materials sectors (3.2%).

Standard RAG: ESG factors can impact corporate financial performance in various ways. Research suggests there is generally a positive relationship between ESG practices and financial outcomes. Studies (58% positive correlation) show that companies with strong ESG performance often experience benefits like improved risk management, enhanced reputation, and increased operational efficiency. Environmental factors appear to have the strongest impact (12.9% higher ROA).

P-RAG: ESG factors increasingly influence corporate financial performance through multiple pathways. Research indicates a generally positive relationship between strong ESG practices and financial outcomes. A meta-analysis from NYU Stern found that 58% of studies showed positive ESG-financial performance relationships, while only 8% were negative. Companies embracing robust ESG frameworks can benefit from improved stakeholder relations, risk mitigation, and market opportunities.

GuarantRAG (Ours): Environmental, social, and governance (ESG) factors demonstrate significant impacts on corporate financial performance through multiple mechanisms. According to a comprehensive 2021 meta-analysis by NYU Stern Center for Sustainable Business examining 1,141 peer-reviewed studies, 58% of corporate studies found positive ESG-financial performance relationships, with merely 8% showing negative correlations. Environmental factors demonstrated the strongest positive impact, with environmental management systems linked to a 12.9% higher ROA. Social practices boosted ROE by 7.5%, while governance improvements delivered a 5.8% increase in Tobin's Q ratio. These financial benefits likely stem from reduced regulatory and reputational risks, operational efficiencies, and enhanced stakeholder relationships. Importantly, ESG impact varies substantially by industry—consumer goods and financial services saw greater benefits (11.5% and 10.2% increased ROE respectively) compared to energy and materials sectors (3.2%). This suggests that ESG integration strategies should be tailored to industry-specific materiality factors rather than applying universal approaches.

Table 24: Case study on ESG factors and corporate financial performance.

Query: What specific advancements has neuromorphic computing made in 2023?

Retrieved Document: In 2023, neuromorphic computing achieved several breakthroughs. Intel's Loihi 2 neuromorphic chip demonstrated real-time learning with 35x lower energy consumption than GPUs on complex robotics tasks. SynSense unveiled the DynapCNN, integrating 5.12 million spiking neurons with 40x power efficiency over traditional chips for edge AI. IBM's TrueNorth system scaled to 1.4 billion neurons and performed unsupervised feature extraction with 96.4% accuracy on medical imaging datasets. Stanford researchers developed memristor arrays with 1024 crossbar junctions achieving 8-bit precision, a 4x improvement over 2022 models.

Standard RAG: Neuromorphic computing has made several advances in 2023. These include improvements in energy efficiency (Intel's Loihi 2 uses 35x less energy than GPUs) and performance. There have also been developments in hardware architecture (SynSense's DynapCNN) and algorithm implementation. The field continues to progress toward more brain-like computing.

P-RAG: Neuromorphic computing, which mimics brain-like neural structures in hardware, has seen significant progress in 2023. Key advancements include improvements in chip architecture, energy efficiency, and scaling capabilities. Notable developments include Intel's Loihi 2 chip and IBM's TrueNorth system scaling to 1.4 billion neurons. These innovations are bringing us closer to efficient, brain-inspired computing systems.

GuarantRAG (Ours): Neuromorphic computing, which implements brain-inspired neural architectures in hardware, has achieved remarkable advancements in 2023 across multiple dimensions. Intel's Loihi 2 neuromorphic chip demonstrated significant breakthroughs in real-time learning while consuming 35x less energy than GPUs when applied to complex robotics tasks. SynSense made substantial contributions with their DynapCNN, which integrates 5.12 million spiking neurons while delivering 40x better power efficiency compared to traditional computing chips for edge AI applications. These efficiency gains are critical for enabling AI deployment in resource-constrained environments. IBM's TrueNorth system represented another significant milestone, scaling to 1.4 billion neurons and achieving 96.4% accuracy on unsupervised feature extraction for medical imaging datasets. At the materials science frontier, Stanford researchers developed advanced memristor arrays with 1024 crossbar junctions capable of 8-bit precision, representing a 4x improvement over 2022 models. Together, these advancements are accelerating neuromorphic computing toward practical applications in robotics, edge intelligence, and medical diagnostics while maintaining the energy efficiency advantages that make this paradigm promising.

Table 25: Case study on 2023 neuromorphic computing advancements.