ASTUTE RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG), while effective in integrating external knowledge to enhance large language models (LLMs), can be undermined by *imperfect* retrieval, which may introduce irrelevant, misleading, or even malicious information. Despite its importance, previous studies have rarely explored the behavior of RAG with errors from imperfect retrieval, and how potential conflicts arise between the LLMs' internal knowledge and external sources. We show that imperfect retrieval augmentation might be inevitable and quite harmful, through controlled analysis under realistic conditions. *Knowledge conflicts* between LLM-internal and external knowledge from retrieval is a bottleneck to overcome in the post-retrieval stage of RAG. To render LLMs resilient to imperfect retrieval, we propose ASTUTE RAG, a novel RAG approach that *adaptively* elicits essential information from LLMs' internal knowledge, *iteratively* consolidates internal and external knowledge with *source-awareness*, and finalizes the answer according to information reliability. Our experiments with Gemini and Claude demonstrate that ASTUTE RAG significantly outperforms previous robustness-enhanced RAG methods. Notably, ASTUTE RAG is the only approach that matches or exceeds the performance of LLMs without RAG under worst-case scenarios. ASTUTE RAG effectively resolves knowledge conflicts, improving the reliability and trustworthiness of RAG systems.

1 Introduction

Retrieval-augmented generation (RAG) is commonly used for large language models (LLMs) to tackle knowledge-intensive tasks (Guu et al., 2020; Lewis et al., 2020). Prior works mainly leverage RAG to address the inherent knowledge limitations of LLMs, effectively integrating missing information and grounding to reliable sources. However, recent research has highlighted a significant drawback that RAG might rely on *imperfect retrieval*, including irrelevant, misleading, or even malicious information (Fig. 1), which eventually leads to inaccurate LLM responses (Chen et al., 2024a; Zou et al., 2024). Moreover, recent studies have shown that retrieval augmentation can confuse LLMs when retrieved passages are *conflicting* with LLMs' parametric knowledge (Tan et al., 2024; Xie et al., 2024; Jin et al., 2024). These pose significant challenges to the trustworthiness of RAG.

To address imperfect retrieval, earlier work seeks to improve the retrieval approaches, such as dynamic and iterative retrieval (Jiang et al., 2023; Asai et al., 2023; Yan et al., 2024). However, the occurrence of imperfect retrieval is still inevitable, due to corpus quality limitations (Shao et al., 2024), the reliability of retrievers (Dai et al., 2024), and the complexity of queries (Su et al., 2024). Consequently, recent work shifts the focus to the generation stage, seeking to reduce the negative impact of noisy retrieved passages (Xiang et al., 2024; Wei et al., 2024). Another line of research at generation stage, motivated by knowledge conflicts, has explored complementing retrieved passages with LLM-generated passages (Yu et al., 2023a; Zhang et al., 2023) or deactivating RAG when the retrieved passages are of insufficient quality (Xu et al., 2024; Mallen et al., 2023; Jeong et al., 2024).

Despite the previous work on the impact of imperfect retrieval and knowledge conflicts at RAG generation stage, quantitative analyses lack on two crucial real-world aspects: (i) the relation between retrieval quality and occurrence of knowledge conflicts, and (ii) the extent to which retrieved passages and LLMs' parametric knowledge can correct each other. Method-wise, existing approaches for mitigating RAG failures caused by imperfect retrieval and knowledge conflicts have not yet yielded a training-free method capable of *explicitly* analyzing conflicting knowledge across various internal

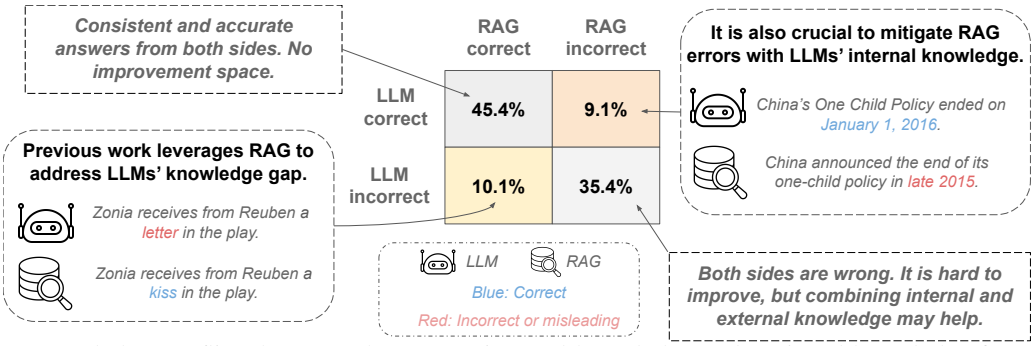

Figure 1: Knowledge conflicts between the LLMs' internal knowledge and retrieved knowledge from external sources. We report the overall results with Claude under the setting in Sec. 5.1.

and external sources, and achieving *worst-case* robustness for black-box LLMs.

In this paper, we first conduct comprehensive analyses to investigate the relation between imperfect retrieval and knowledge conflicts, and examine the frequency of external and LLMs' internal knowledge mutually correcting each other (Sec. 3). On a diverse range of general, domain-specific, and long-tail questions from NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), BioASQ (Tsatsaronis et al., 2015), and PopQA (Mallen et al., 2023), we observe that imperfect retrieval is widespread even with an adept real-world search engine, leading to the impeded performance of RAG.[1] Retrieval precision is tightly correlated with the knowledge conflict rate. Mutual correction between the LLM's knowledge and external knowledge is crucial for recovering from RAG failures. Our findings underscore the potential severity of imperfect retrieval in real-world RAG and highlight the widespread existence of knowledge conflicts as the bottleneck.

We propose ASTUTE RAG, a novel RAG approach designed for resilience to imperfect retrieval augmentation, while preserving RAG grounding effect when retrieval is reliable (Sec. 4). ASTUTE RAG effectively differentiates between consistent and conflicting information from the LLM's internal knowledge and the externally retrieved passages, assesses their reliability, and ensures proper integration of trustworthy information. ASTUTE RAG first adaptively elicits LLMs' knowledge and then conducts source-aware knowledge consolidation. The desiderata is combining consistent information, identifying conflicting information, and filtering out irrelevant information. Finally, ASTUTE RAG proposes answers based on consistent information and compares them to determine the final answer. Our experiments with various LLMs (Claude, Gemini and Mistral), demonstrate superior performance of ASTUTE RAG compared to previous RAG approaches designed for robustness (Sec. 5). ASTUTE RAG consistently outperforms baselines across different retrieval quality levels. Notably, ASTUTE RAG is the only RAG method that achieves performance comparable to or even surpassing retrieval-free mode of LLMs under the worst-case scenario where all retrieved passages are unhelpful. Further analysis reveals the effectiveness of ASTUTE RAG in resolving knowledge conflicts.

In summary, our core contributions are threefold. First, we provide quantitative analyses and novel insights for the connection among imperfect retrieval, knowledge conflicts, and RAG failures under real-world conditions. Second, we propose ASTUTE RAG, which explicitly analyzes LLM-internal and external knowledge in-context, assesses their reliablity, and recovers from RAG failures with black-box access. Third, with experiments with various LLMs and datasets, we demonstrate the effectiveness of ASTUTE RAG in improving robustness and trustworthiness, even in the most challenging scenarios.

2 Related Work

RAG aims to address the inherent knowledge limitation of LLMs with passages retrieved from external sources of information such as private corpora or public knowledge bases (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022). Given the widespread real-world adoption of RAG, including risk-sensitive domains, the negative impact of noisy information within retrieved passages has garnered increasing attention (Cuconasu et al., 2024). Recent work explored enhancing the robustness of RAG systems against noise from various perspectives, including training LLMs with noisy context (Yu et al., 2023b; Yoran et al., 2024; Pan et al., 2024; Fang et al., 2024), training small models to filter out irrelevant passages (Wang et al., 2023b;

[1] such as Google Search with Web as corpus

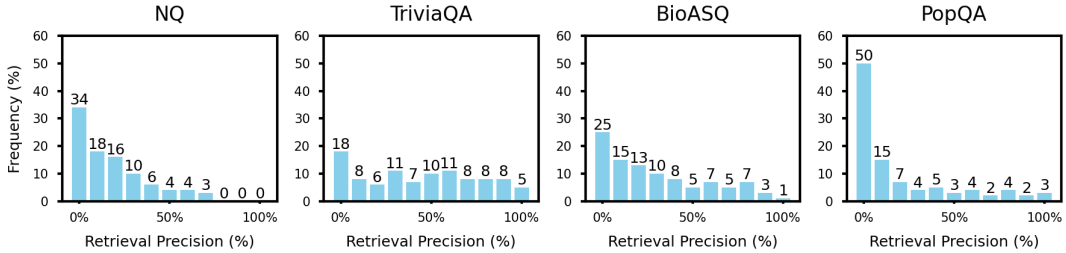

Figure 2: Imperfect retrieval (samples with low retrieval precision) is prevalent in real-world RAG.

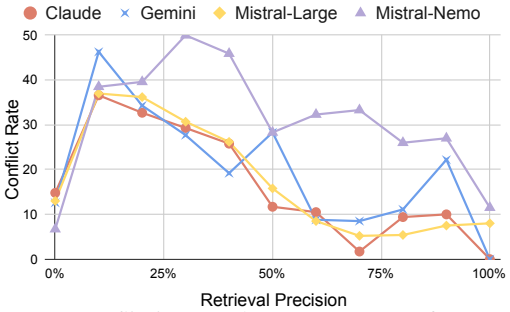

Figure 3: Conflicting rate between answers from LLMs with and without RAG on different retrieval precision.

Xu et al., 2023), passage reranking (Yu et al., 2024; Glass et al., 2022), dynamic and iterative retrieval (Jiang et al., 2023; Asai et al., 2023; Yan et al., 2024), query rewriting (Ma et al., 2023), and speculative drafting (Wang et al., 2024). These focus on distinct modules or stages of RAG systems and are orthogonal to our work.

Our work focuses on enhancing RAG robustness at the post-retrieval stage, after retrieved passages have been provided. On this, RobustRAG (Xiang et al., 2024) aggregates answers from each independent passage to provide certifiable robustness. InstructRAG (Wei et al., 2024) instructs the LLM to provide a rationale connecting the answer with information in passages. MADRA (Wang et al., 2023a) applies multi-agent debate to select helpful evidence. However, these do not explicitly incorporate internal knowledge to recover from RAG failures and therefore might severely suffer when the majority of retrieved passages have issues. For emphasizing internal knowledge of LLMs in RAG, recent work explored using LLM-generated passage as context (Yu et al., 2023a), training models to match generated and retrieved passages (Zhang et al., 2023), adaptively switching between LLMs with and without RAG (Xu et al., 2024; Mallen et al., 2023; Jeong et al., 2024), and combining answers through contrastive decoding (Zhao et al., 2024; Jin et al., 2024). Different from prior work, we provide a systematic framework on connecting imperfect retrieval, knowledge conflicts, and RAG failures. Specifically focusing on the imperfect context setting, our method is training-free and applicable to black-box LLMs, explicitly analyzes internal and external knowledge in-context, and offers broader usability and adaptability.

3 The Pitfall of RAG

To better showcase common real-world challenges and motivate improved methodological designs, we evaluate retrieval quality, the occurrence of knowledge conflicts, their relationship, and the mutual correction between external and internal knowledge using a controlled dataset derived from NQ, TriviaQA, BioASQ, and PopQA, datasets widely used for RAG in prior work (Xiang et al., 2024; Wei et al., 2024; Asai et al., 2023). Different from prior work, our analysis is based on real-world retrieval results with Google Search[2] as the retriever and the Web as the corpus. Overall, we sample 1K instances, each with 10 retrieved passages.

Imperfect retrieval and knowledge conflicts are common and harmful. Our initial observations are consist with prior work. As shown in Fig. 2, the retrieval precision[3] is generally low - roughly 70% retrieved passages do not directly contain true answers, consistent with prior work demonstrating the often imperfect nature of retrieval results (Thakur et al., 2024; Su et al., 2024). With Claude 3.5 Sonnet as the LLM, Fig. 1 shows that 19.2% of the overall data exhibit knowledge conflicts, consistent with prior work demonstrating the prevalence of such conflicts across various scenarios (Pham et al., 2024; Xie et al., 2024; Longpre et al., 2021). Moreover, we observe strong correlations between retrieval precision and RAG performance (Fig. 7) and between the occurrence of knowledge conflicts and RAG performance (Fig. 8), findings consistent with prior work on these respective topics (Chen et al., 2024a; Xie et al., 2024).

Lower retrieval precision increases knowledge conflicts in general. As shown in Fig. 3, most advanced LLMs exhibit the highest conflict rates

[2] https://developers.google.com/custom-search/v1/overview

[3] Ratio of passages directly contain true answers.

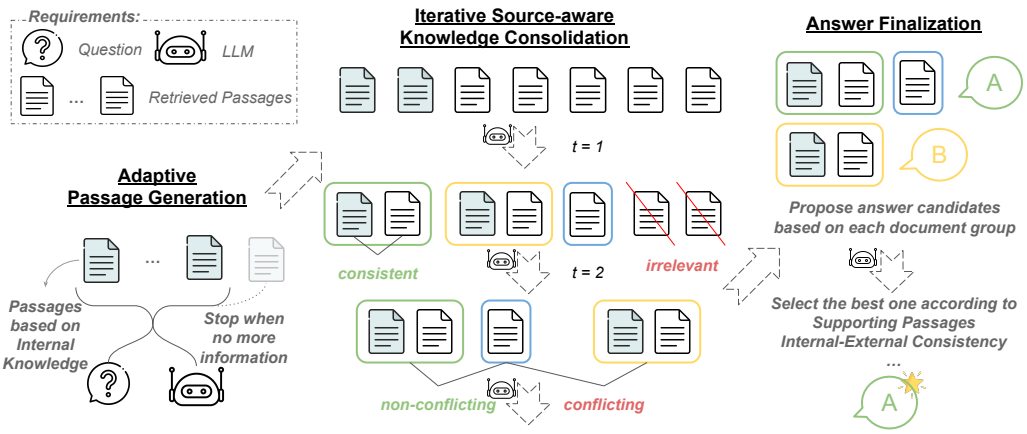

Figure 4: Overview of the ASTUTE RAG framework. ASTUTE RAG is designed to better combine the information from the external sources (e.g. web, domain-specific corpora) and internal knowledge of the LLMs by employing a consolidation mechanism to address the conflicts, which eventually leads to superior generation quality.

when retrieval precision is as low as 10%. Subsequently, the conflict rate generally decreases as precision increases, although some fluctuations may occur. This trend is generally applicable to the studied LLMs with different training processes. Notably, when retrieval precision is 0%, conflict rates tend to be significantly lower. This suggests that limited external knowledge for the query results in more irrelevant passages rather than incorrect ones.

Internal and external knowledge can correct each other to a comparable extent. Among the conflicting cases, the internal knowledge is correct on 47.4% of them, while the external knowledge is correct on the remaining 52.6%. These results emphasize the importance of effectively combining the internal and external knowledge to overcome the inherent limitation of relying solely on either source. However, previous work (Tan et al., 2024; Xie et al., 2024; Jin et al., 2024) shows that LLMs often select knowledge based on unreliable shortcuts, so simply presenting LLM-generated passages in the context may not help.

4 ASTUTE RAG

We first provide an overview of ASTUTE RAG (Sec. 4.1). Subsequently, we delve into the three major steps of ASTUTE RAG, including adaptive generation of internal knowledge (Sec. 4.2), source-aware knowledge consolidation (Sec. 4.3), and answer finalization (Sec. 4.4).

4.1 Overview

Our objective is to mitigate the effects of imperfect retrieval augmentation, resolve knowledge conflicts between the LLM's internal knowledge and external sources (such as custom/public corpora and knowledge bases), and ultimately produce more accurate and reliable responses from LLMs. Given a set of retrieved passages from external sources $E = [e_1, \ldots, e_n]$, a pre-trained LLM $\mathcal{M}$ (accessible through prediction-only APIs, encompassing commercial black-box ones), and a query q, the task is to generate the corresponding correct answer a^*. Notably, this setting is orthogonal to prior work on improving the retriever, training LLMs, or conducting adaptive retrieval, which are mainly preliminary steps.

ASTUTE RAG is designed to better leverage collective knowledge from both internal knowledge of LLMs and external corpus, for more reliable responses. As shown in Fig. 4 and Alg. 1, ASTUTE RAG starts from acquiring the most accurate, relevant, and thorough passage set from the LLMs' internal knowledge. Then, internal and external knowledge are consolidated in an iterative way, by comparing the generated and retrieved passages. Finally, the reliability of conflicting information is compared and the final output is generated according to the most reliable knowledge.

4.2 Adaptive Generation of Internal Knowledge

In the first step, we elicit internal knowledge from LLMs. This LLM-internal knowledge, reflecting the consensus from extensive pre-training and instruction-tuning data, can supplement any missing information from the limited set of retrieved passages and enable mutual confirmation between LLM-internal and external knowledge. This is especially valuable when the majority of retrieved passages might be irrelevant or misleading. Specifically, we prompt LLMs to generate passages based on the given question q, following Yu et al. (2023a). While Yu et al. (2023a) primarily focused on generating diverse internal passages, we emphasize

Algorithm 1 ASTUTE RAG

Require: Query q, Retrieved Passages $E = [e_1, \ldots, e_n]$, Large Language Model $\mathcal{M}$, Number of Iteration t, Max Number of Generated Passages $\hat{m}$, Prompt Templates $p_{gen}, p_{con}, p_{ans}$

Adaptively generate passages: $I \leftarrow \mathcal{M}(p_{gen}, q, \hat{m})$ ▷ Sec. 4.2
Combine internal and external passages: $D_0 \leftarrow E \oplus I$
Assign passage sources: $S_0 \leftarrow [\mathbb{1}_{\{d \in E\}} \text{for } d \text{ in } D_0]$
if $t > 1$ **then**
  for $j = 1, \ldots, t-1$ **do** ▷ Sec. 4.3
    Consolidate knowledge: $\langle D_{j+1}, S_{j+1} \rangle \leftarrow \mathcal{M}(p_{con}, q, \langle D_0, S_0 \rangle, \langle D_j, S_j \rangle)$
  end for
  Finally consolidate and answer: $a \leftarrow \mathcal{M}(p_{ans}, q, \langle D_0, S_0 \rangle, \langle D_{t-1}, S_{t-1} \rangle)$ ▷ Sec. 4.4
else
  Consolidate knowledge and finalize the answer: $a \leftarrow \mathcal{M}(p_{ans}, q, \langle D_0, S_0 \rangle)$
end if
return a

the importance of reliability and trustworthiness of generated passages. To achieve this goal, we enhance the original method with *constitutional principles* and *adaptive generation*.

Inspired by Bai et al. (2022), we provide **constitutional principles** indicating the desired properties of internal passages in the prompt p_{gen} (see Appx. A for details) to guide their generation, emphasizing that the generated passages should be accurate, relevant, and hallucination-free. Moreover, we allow the LLM to perform **adaptive generation** of passages in its internal knowledge. The LLM can decide how many passages to generate by itself. Rather generating a fix number of passages, we request the LLM to generate at most $\hat{m}$ passages, each covering distinct information, and to directly indicate if no more reliable information is available. This adaptive approach allows the LLM to generate fewer passages (or even no passages at all) when the useful information within internal knowledge is limited and more passages when there are multiple feasible answers in the internal knowledge. In this step, the LLM generates $m \leq \hat{m}$ passages based on its internal knowledge:

$$I = [i_1, \ldots i_m] = \mathcal{M}(p_{gen}, q, \hat{m}).$$

4.3 Iterative Source-aware Knowledge Consolidation

In the second step, we employ the LLM to explicitly consolidate information from both passages generated from its internal knowledge and passages retrieved from external sources. Initially, we combine passages from both internal and external knowledge sources $D_0 = E \oplus I$.

We additionally ensure **source-awareness** by providing the source of each passage to LLMs when consolidating knowledge. The source information (internal or external, such as a website) is helpful in assessing the reliability of passages. Here, we provide the passage source as $S_0 = [\mathbb{1}_{\{d \in E\}} \text{for } d \text{ in } D_0]$. To consolidate knowledge, we prompt the LLM (with p_{con} in Appx. A) to identify consistent information across passages, detect conflicting information between each group of consistent passages, and filter out irrelevant information. This step would regroup the unreliable knowledge in input passages into fewer refined passages. The regrouped passages also attribute their source to the corresponding input passages:

$$\langle D_{j+1}, S_{j+1} \rangle = \mathcal{M}(p_{con}, q, \langle D_0, S_0 \rangle, \langle D_j, S_j \rangle).$$

We find that this is especially helpful in comparing the reliability of conflicting knowledge and addressing knowledge conflicts. This knowledge consolidation process can run **iteratively** for t times to improve better utilization of the retrieved context.

4.4 Answer Finalization

In the last step, we prompt the LLM (with p_{ans} in Appx. A) to generate one answer based on each group of passages ($\langle D_t, S_t \rangle$), and then compare their reliability and select the most reliable one as the final answer. This comparison allows the LLM to comprehensively consider knowledge source, cross-source confirmation, frequency, and information thoroughness when making the final decision. Notably, this step can be merged into the last knowledge consolidation step to reduce the inference complexity (the amount of prediction API calls) using a combined prompt:

$$a = \mathcal{M}(p_{ans}, q, \langle D_0, S_0 \rangle, \langle D_t, S_t \rangle).$$

When $t = 1$, the initial passages will be input to the model directly for knowledge consolidation and subsequent answering: $a = \mathcal{M}(p_{ans}, q, \langle D_0, S_0 \rangle)$.

Method	NQ	TriviaQA	BioASQ	PopQA	Overall	NQ	TriviaQA	BioASQ	PopQA	Overall
	Claude 3.5 Sonnet (20240620)					*Gemini 1.5 Pro (002)*				
No RAG	47.1	82.0	50.4	29.8	54.5	44.8	80.2	45.8	25.3	51.3
RAG	44.4	76.7	58.0	36.0	55.5	42.7	76.0	55.2	33.7	53.7
USC (Chen et al., 2024b)	48.1	80.2	**61.5**	37.6	58.7	46.4	76.7	**58.4**	37.6	56.4
GenRead (Yu et al., 2023a)	42.0	74.2	57.0	34.3	53.6	45.1	77.4	54.9	34.3	54.7
RobustRAG (Xiang et al., 2024)	47.8	78.1	56.3	37.1	56.5	34.2	67.5	44.1	32.0	45.6[4]
InstructRAG (Wei et al., 2024)	47.1	83.0	58.0	41.0	58.8	46.8	80.6	54.9	34.8	56.1
Self-Route (Xu et al., 2024)	47.5	78.8	59.1	41.0	58.1	47.5	79.9	58.0	38.2	57.6
ASTUTE RAG	**52.2**	**84.1**	60.1	**44.4**	**61.7**	**50.2**	**81.6**	58.0	**40.5**	**59.2**
	Mistral-Large (2407), 128B					*Mistral-Nemo (2407), 12B*				
No RAG	46.8	79.5	43.7	24.7	51.1	29.8	67.8	34.3	23.0	40.2
RAG	43.1	77.4	55.9	36.0	54.7	39.3	66.8	49.0	32.6	48.3
USC (Chen et al., 2024b)	**51.2**	80.9	**61.5**	36.0	59.5	29.5	66.1	36.0	20.2	39.6
GenRead (Yu et al., 2023a)	40.7	73.1	55.6	35.4	52.7	38.6	68.9	48.3	**33.7**	48.7
RobustRAG (Xiang et al., 2024)	42.7	77.7	50.4	34.8	53.0	35.6	71.7	44.1	27.5	46.4
InstructRAG (Wei et al., 2024)	45.4	80.6	57.3	36.5	56.7	38.3	61.8	50.4	23.6	45.5
Self-Route (Xu et al., 2024)	45.4	77.7	57.3	38.2	56.2	41.4	73.5	**51.8**	30.9	51.2
ASTUTE RAG	50.2	**82.7**	58.4	**42.1**	**59.9**	**42.7**	**73.9**	49.3	32.6	**51.3**

Table 1: Main results on Claude 3.5 Sonnet, Gemini 1.5 Pro, Mistral-Large, and Mistral-Nemo under zero-shot setting, showing the accuracy of benchmarked alternatives vs. ASTUTE RAG. Best scores are in bold. Note that USC consumes approximately three times more tokens than other RAG methods, and is not directly comparable.

5 Experiments

We evaluate the effectiveness of ASTUTE RAG on overcoming imperfect retrieval augmentation and addressing knowledge conflicts. In this section, we introduce the experiment setting (Sec. 5.1), compare the performance of ASTUTE RAG with various baselines on diverse datasets (Sec. 5.2), and provide in-depth analyses (Sec. 5.3).

5.1 Experimental Settings

Datasets and metrics. We consider datasets encompass general questions, domain-specific questions, long-tail questions, as well as both short-form and long-form formats, following prior work (Xiang et al., 2024; Wei et al., 2024). On NQ, TriviaQA, BioASQ, and PopQA, we provide 10 passages collected with Google Search from the Web for each instance. For long-form QA, we use ASQA (Stelmakh et al., 2022). We also evaluate on RGB (Chen et al., 2024a). We choose the English subset (refined version) focusing on noise robustness. For each instance, we select five top negative passages to form a worst-case scenario. Following prior work, we report the accuracy by string match. More details are in Appx. B.

Models and General Settings. We conduct experiments on advanced proprietary and open-source LLMs of different scales, including Claude 3.5 Sonnet (`claude-3-5-sonnet@20240620`),[5] Gemini 1.5 Pro (`gemini-1.5-pro-002`),[6] Mistral-Large (`128B; version 2407`), and Mistral-Nemo (`12B; version 2407`). The generation temperature is set to 0 and the maximum output tokens is set to 1,024. All experiments are under the zero-shot setting for controlled evaluation.

Baselines. We compare ASTUTE RAG with various RAG methods designed for enhanced robustness. *USC* (Chen et al., 2024b) is a self-consistency method that samples multiple LLM responses and aggregates the answers. It provides a reference of naive improvements using additional API calls. *Genread* (Yu et al., 2023a) augments retrieved passages with LLM-generated passages without explicit consolidation process. *RobustRAG* (Xiang et al., 2024) aggregates answers from independent passages to provide certifiable robustness. We use the best-performing keyword aggregation variant. *InstructRAG* (Wei et al., 2024) instructs the LLM to provide a rationale connecting the answer with information in passages. For a fair comparison, no training is applied. *Self-Route* (Xu et al., 2024) adaptively switches between LLMs with and without RAG.[7] It provides a reference of switching between LLMs' internal and external knowledge.

[4] We observe a high refusal rate in RobustRAG for Gemini.

[5] https://www.anthropic.com/news/claude-3-5-sonnet

[6] https://deepmind.google/technologies/gemini/pro/

[7] The original Self-Route switches between RAG and long-context LLMs, while our implementation switches between RAG and No RAG according to our problem formulation.

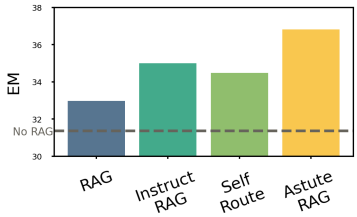

Figure 5: Performance on ASQA.

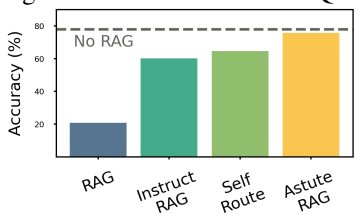

Figure 6: Worst-case performance of Claude on RGB. ASTUTE RAG reaches a performance close to No RAG, while other RAG systems are far behind.

Implementation Details. The prompt templates for ASTUTE RAG can be found in Appx. A. By default, we set $t = 1$ and $\hat{m} = 1$ to limit the number of additional tokens used. Results with larger t and $\hat{m}$ are discussed in Sec. 5.3.

5.2 Main Results

Performance under real-world retrieval. Tab. 1 presents the results with real-world retrieval augmentation of various LLMs. We find that retrieved passages might not always bring benefits – on NQ and TriviaQA, RAG performance lags behind No RAG for advanced LLMs. We attribute this questions being covered by the LLM's internal knowledge and the noise in retrieval results misleading the LLM. In contrast, on BioASQ and PopQA, which focus on domain-specific and 'long-tail' questions, RAG significantly improves the LLM performance. Due to imperfect retrieval augmentation, however, the absolute performance still remains to be unsatisfactory. Among all baselines, no single method consistently outperforms others across all datasets and LLMs. This observation highlights these baselines being tailored to distinct settings and not being universally applicable. Overall, InstructRAG and Self-Route demonstrate relatively superior performance among other alternatives. ASTUTE RAG consistently outperforms baselines across all LLMs in terms of overall accuracy. The relative improvement compared to the best baseline is 6.85% for Claude and 4.13% for Gemini, with the improvements in domain-specific questions being much higher. These highlight the effectiveness of ASTUTE RAG in overcoming imperfect retrieval augmentation and knowledge conflicts. Additionally, we observe consistent improvements on the open-source Mistral models. The results demonstrate that ASTUTE RAG generalizes well to LLMs of smaller sizes.

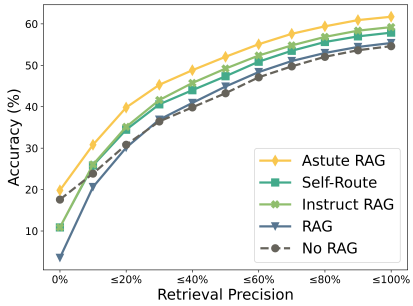

Figure 7: Performance across different retrieval precision buckets. ASTUTE RAG is consistently better.

Performance on long-form QA. We conduct additional experiments on the long-form QA dataset, ASQA. Fig. 5 demonstrates that ASTUTE RAG consistently achieves significant improvements, reinforcing its effectiveness across diverse scenarios.

Worst-case performance on RGB. Fig. 6 presents the results under the worst-case setting on RGB where all retrieved documents are negative, to demonstrate robustness. The performance gap between RAG and No RAG exceeds 50 points, highlighting the detrimental impact of imperfect retrieval results and emphasizing the importance of providing robust safeguards against worst-case scenarios. While the baseline RAG methods outperform the original RAG, they still obviously fall behind 'No RAG'. ASTUTE RAG is the only RAG method that reaches a performance close to 'No RAG', further supporting its effectiveness in addressing imperfect retrieval augmentation.

5.3 Analyses

We conduct in-depth analyses using Claude following the setting of Tab. 1.

The impact of retrieval precision. As shown in Fig. 7, ASTUTE RAG achieves consistently better performance across different retrieval precision regimes, indicating its effectiveness in improving RAG trustworthiness in broad scenarios. Notably, ASTUTE RAG does not sacrifice performance gain under high retrieval quality in exchange for improvement under low retrieval quality. When the retrieval quality is extremely low (close to zero precision), all other RAG variants underperform the 'No RAG' baseline, except for ASTUTE RAG.

Addressing knowledge conflicts. We split our collected data into three subset according to the answers with and without RAG: the answers from two can be (i) both correct, (ii) both incorrect, or (iii) conflicting with one being correct. The results

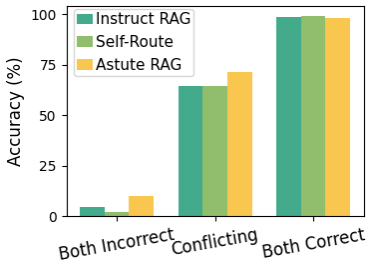

Figure 8: Performance on conflicting and consistent instances between No RAG and RAG.

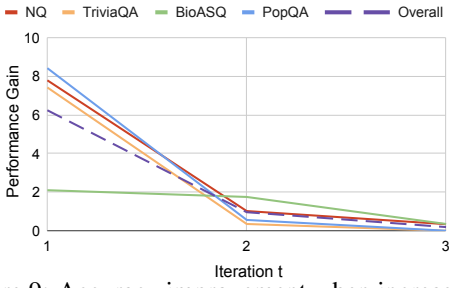

Figure 9: Accuracy improvement when increasing t.

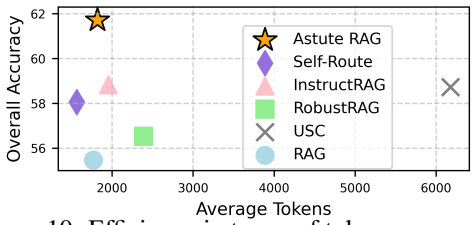

Figure 10: Efficiency in terms of tokens consumed.

are shown in Fig. 8. On the conflicting subset, ASTUTE RAG successfully chooses the correct answer in approximately 80% of cases, being the most effective one in addressing knowledge conflicts. Notably, ASTUTE RAG even brings performance improvement on the subset where neither internal nor external knowledge alone leads to the correct answer. This indicates that ASTUTE RAG can effectively combine partially-correct information from LLM-internal and external knowledge.

Benefits of more consolidation iteration. For efficiency, we employ a single iteration of knowledge consolidation in our main experiments. However, incorporating multiple iterations has the potential to further enhance model performance as shown in Fig. 9. The magnitude of this improvement diminishes as t increases, indicating that the knowledge has been better presented and less improvement space left after each iteration.

Effectiveness of adaptive generation. The results in Tab. 2 illustrate the model's performance when varying the maximum number of passages generated. The design of adaptive generation has been effectively reflected, as the number of generated passages is dynamically adjusted leading to $m < \hat{m}$. Notably, the number of generated passages can be controlled by $\hat{m}$, and results show that the system does not generate passages excessively.

Efficiency in tokens consumed and API calls. As a proxy to overall prediction cost and latency, we present the average number of tokens and API calls used per instance in Fig. 10 and Fig. 11. ASTUTE RAG incurs only a marginal cost increase, <5%, while delivering substantial improvement, >11%, compared to the RAG baseline.

Accuracy of intermediate steps. To investigate the performance of intermediate steps, including knowledge consolidation and confidence assignment, we use LLM-as-a-judge with the instruction in Appx. A. Our experimental results show that the accuracy for knowledge consolidation is 98.2%, and for confidence assignment, it is 95.0%. These results demonstrate the effectiveness of the proposed framework in the intermediate stages.

Influence of passage ordering. We apply different ordering strategies (Alessio et al., 2024), on RAG and ASTUTE RAG. As shown in Tab. 3, we find that the improvement with ASTUTE RAG is significantly larger than the gap between different ordering strategies. Moreover, the consolidation process makes ASTUTE RAG less sensitive to it.

Qualitative examples. In Fig. 12, we present two representative examples showing the intermediate outputs of ASTUTE RAG. In the first example, LLM without RAG generates a wrong answer, while RAG returns a correct answer. ASTUTE RAG successfully identified the incorrect information in its generated passage and an external passage, avoiding confirmation bias (Tan et al., 2024). In the second example, LLM is correct but RAG is incorrect due to imperfect retrieval. ASTUTE RAG detected the correct answer from imperfect context leveraging internal knowledge.

6 Conclusion

We investigate the impact of imperfect retrieval on the performance of RAG systems and identify knowledge conflicts as a key challenge. To address this, we introduce ASTUTE RAG, a novel approach that leverages the internal knowledge of LLMs and iteratively refines the generated responses by consolidating internal and external knowledge in a source way. We demonstrate the effectiveness of ASTUTE RAG in mitigating the negative effects of imperfect retrieval and improving the robustness of RAG, particularly in challenging scenarios with unreliable external sources.

Limitations

ASTUTE RAG's effectiveness hinges on the capabilities of advanced LLMs with strong instruction-following and reasoning abilities, hence potentially more limited applicability with less sophisticated LLMs. As an important future direction, extending the experimental setup to include longer inputs would be important, where the challenges of imperfect retrieval and knowledge conflicts may be even more pronounced.

References

Marco Alessio, Guglielmo Faggioli, Nicola Ferro, Franco Maria Nardini, Raffaele Perego, et al. 2024. Improving rag systems via sentence clustering and reordering. In *RAG@ SIGIR 2024 workshop: The Information Retrieval's Role in RAG Systems, ACM*, pages 1–10.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024b. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. *arXiv preprint arXiv:2404.11457*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *arXiv preprint arXiv:2405.20978*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

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.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching llms credibility-aware generation. *arXiv preprint arXiv:2404.06809*.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. *arXiv preprint arXiv:2410.15737*.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. *arXiv preprint arXiv:2407.12854*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv preprint arXiv:2401.11911*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2024. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2023a. Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates. *arXiv preprint arXiv:2312.04854*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, et al. 2024. Speculative rag: Enhancing retrieval augmented generation through drafting. *arXiv preprint arXiv:2407.08223*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation with explicit denoising. *arXiv preprint arXiv:2406.13629*.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, et al. 2024. Crag–comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. 2024. Compact: Compressing retrieved documents actively for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21424–21439.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *arXiv preprint arXiv:2407.02485*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain qa. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. 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 4225–4237.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*.

A Prompt Template for ASTUTE RAG

Prompt for Adaptive Passage Generation (p_{gen})

Generate a document that provides accurate and relevant information to answer the given question. If the information is unclear or uncertain, explicitly state 'I don't know' to avoid any hallucinations.

Question: {question} Document:

Prompt for Iterative Knowledge Consolidation (p_{con})

Task: Consolidate information from both your own memorized documents and externally retrieved documents in response to the given question.

* For documents that provide consistent information, cluster them together and summarize the key details into a single, concise document.
* For documents with conflicting information, separate them into distinct documents, ensuring each captures the unique perspective or data.
* Exclude any information irrelevant to the query.
For each new document created, clearly indicate:
* Whether the source was from memory or an external retrieval.
* The original document numbers for transparency.

Initial Context: {context}
Last Context: {context}
Question: {question}
New Context:

Prompt for Knowledge Consolidation and Answer Finalization (p_{ans})

Task: Answer a given question using the consolidated information from both your own memorized documents and externally retrieved documents.

Step 1: Consolidate information
* For documents that provide consistent information, cluster them together and summarize the key details into a single, concise document.
* For documents with conflicting information, separate them into distinct documents, ensuring each captures the unique perspective or data.
* Exclude any information irrelevant to the query.
For each new document created, clearly indicate:
* Whether the source was from memory or an external retrieval.
* The original document numbers for transparency.

Step 2: Propose Answers and Assign Confidence
For each group of documents, propose a possible answer and assign a confidence score based on the credibility and agreement of the information.

Step 3: Select the Final Answer
After evaluating all groups, select the most accurate and well-supported answer.
Highlight your exact answer within <ANSWER> your answer </ANSWER>.

Initial Context: {context_init}
[Consolidated Context: {context}] # optional
Question: {question}
Answer:

Prompt for Intermediate Step Evaluation

Task: You are provided with the following:
1. A question.
2. The correct answer.
3. The input context.
4. The model's response, which contains:
- Consolidated context.
- Confidence scores for candidate answers.
Your task is to:
- Evaluate the **quality of the consolidated context** in the model's response and provide a label: '<consolidation> correct </consolidation>' or '<consolidation> incorrect </consolidation>'. This evaluation is only about whether the consolidation is correct given the input context.
- Evaluate the **accuracy of the confidence score** (whether it aligns with the confidence of the supporting context) and provide a label: '<confidence> correct </confidence>' or '<confidence> incorrect </confidence>'. The evaluation is only based on the consolidated context.
Note that correct consolidation and confidence do not necessarily indicate the correct answer.
Question: {query}
Correct Answer: {answer}
Input Context: {input}
Model Response: {response}
Evaluation:

B Data Collection

Encompassing a *diverse* range of *natural* questions, our benchmark consists of *realistic* retrieval results with Google Search[8] as the retriever and the Web as the corpus. Notably, we do not select questions or annotate answers based on the retrieval results. This setting allows us to analyze the severity of imperfect retrieval in real-world RAG. It distinguishes our benchmark from previous ones that employ synthetic retrieval corruptions or that unintentionally reduce the frequency of imperfect retrieval with biased construction protocols (Chen et al., 2024a; Yang et al., 2024). Overall, our benchmark contains 1,042 short-form question-answer pairs, each paired with 10 retrieved passages. When collecting the passages, we retrieve the top 30 results and select the first 10 accessible websites. From each retrieved website, we extract the paragraph corresponding to the snippet provided in the search results as the retrieved passage. Retrieved results might contain natural noise with irrelevant or misleading information. We do not consider enhancements to the retrieval side, such as query rewriting, as such enhancements are typically already incorporated into commercial information retrieval systems. All of these datasets are short-form QA. Following previous work (Xiang et al., 2024; Wei et al., 2024; Mallen et al., 2023), a model response is considered correct if it contains the ground-truth answer. To enhance evaluation reliability, we prompt LLMs to enclose the exact answer within special tokens, extracting them as the final responses.

Question-answer pairs. We consider question-answer pairs from four datasets of different properties spanning across general questions, domain-specific questions, and long-tail questions. NQ (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) are two widely-studied question-answering (QA) datasets in general domains. BioASQ (Tsatsaronis et al., 2015) is from biomedical domain that has demonstrated significant benefits from RAG when general-purpose LLMs are considered. PopQA (Mallen et al., 2023) focuses on long-tail knowledge and has been shown to be challenging for even advanced LLMs to solve without external knowledge. All these datasets contain questions with short-form answers and most of them list all valid answer variants. This format can support automatic verification of answer appearance in retrieved passages and model responses, leading to more precise evaluations.

Retrieval process. For each question in our benchmark, we query Google Search to retrieve the top 30 results and select the first 10 accessible websites. From each retrieved website, we extract the paragraph corresponding to the snippet provided in Google Search results as the retrieved passage. We do not consider enhancements to the retrieval side, such as query rewriting, as such enhancements are typically already incorporated into commercial information retrieval systems.

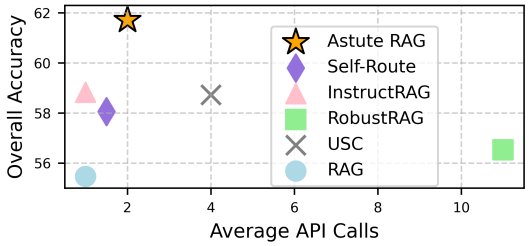

Figure 11: Efficiency in terms of API calls.

C Comparison with Evidence Filtering

Evidence filtering (Wang et al., 2023b; Yoon et al., 2024) is also a related direction. We further conduct experiments comparing our method with CompAct (Yoon et al., 2024). The results in Tab. 4 and Tab. 5 show that evidence filtering is ineffective in handling the challenges of imperfect context and knowledge conflicts. Notably, it even performs worse than the No RAG and RAG baselines in this context. The primary reason for this underperformance lies in the limitations of context compression. It struggles to effectively identify incorrect information when there are conflicts in context and often filters out or reduces the appearance of helpful information in the process. This reinforces the importance of our approach, which does not rely solely on filtering but instead integrates both internal and external knowledge while handling conflicts in a more nuanced manner.

[8] https://developers.google.com/custom-search/v1/overview

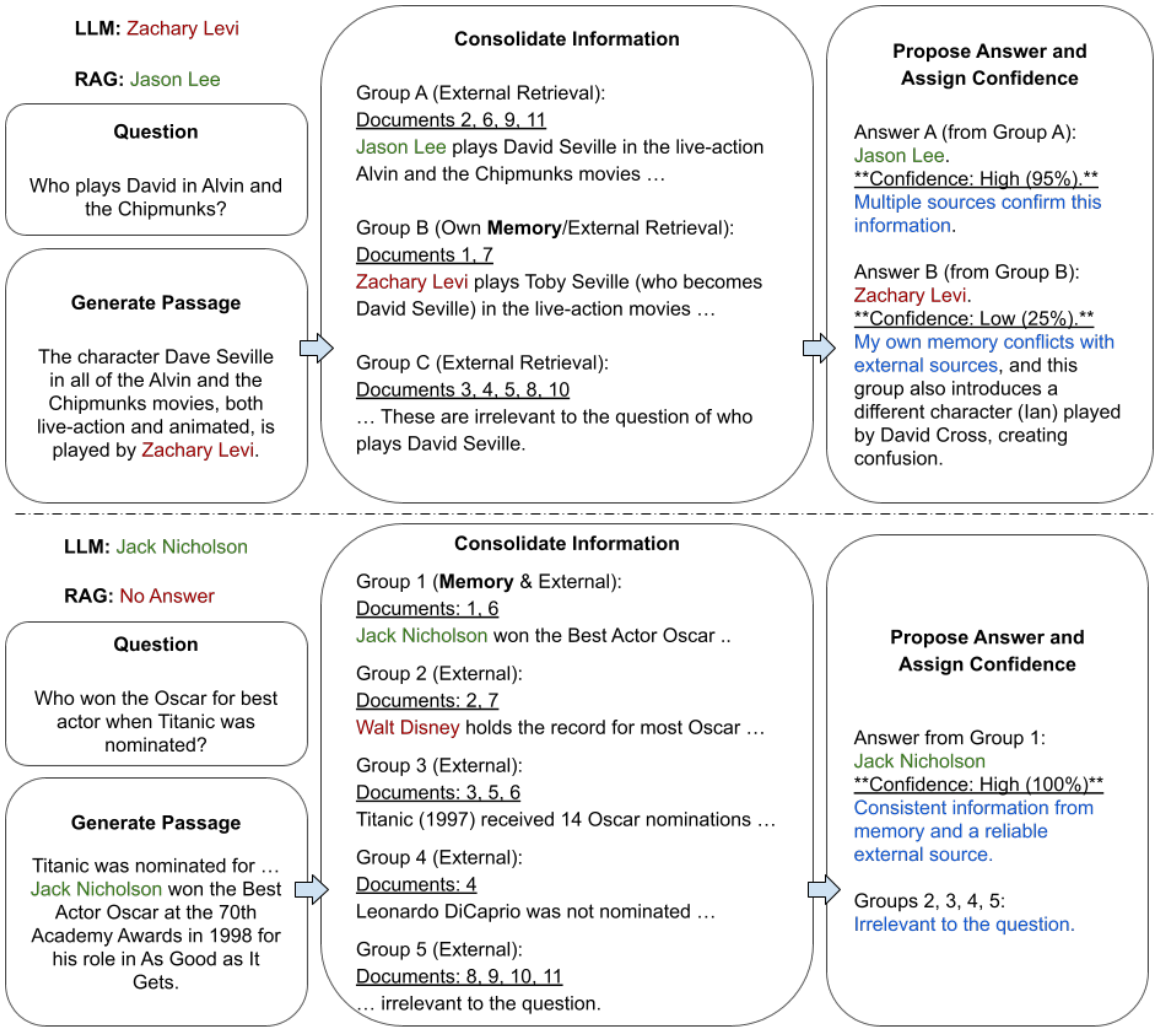

Figure 12: Qualitative examples. *Top:* ASTUTE RAG identified the error in internal knowledge (i.e., generated passage) by confirming with external sources. *Bottom:* ASTUTE RAG detected the correct answer from imperfect retrieval by checking with its internal knowledge. Standard RAG does not provide an answer because the retrieved passages are too noisy.

	NQ	TriviaQA	BioASQ	PopQA	Overall	m
$\hat{m}$=1	52.20	84.10	60.14	44.38	61.71	0.69
$\hat{m}$=2	52.20	85.16	60.84	43.26	62.00	1.24

Table 2: Performance and averge number of generaed passages using different $\hat{m}$.

Method	Ordering Strategy	NQ	TriviaQA	BioASQ	PopQA	Overall
RAG	Random	43.39	76.33	56.99	34.83	54.61
	Ascending	43.05	75.62	57.69	34.83	54.51
	Descending	44.41	76.68	58.04	35.96	55.47
	Ping-pong Descending Top-to-bottom	44.75	77.39	57.69	35.96	55.66
	Ping-pong Descending Bottom-to-top	44.41	75.62	58.04	35.96	55.18
AstuteRAG	Random	51.86	84.81	61.19	41.57	61.61
	Ascending	51.86	85.51	59.79	42.13	61.52
	Descending	52.20	84.10	60.14	44.38	61.71
	Ping-pong Descending Top-to-bottom	52.20	84.45	59.09	43.82	61.42
	Ping-pong Descending Bottom-to-top	51.19	85.16	61.54	43.82	62.00

Table 3: Performance by Ordering Strategies.

Method	NQ	TriviaQA	BioASQ	PopQA	Overall
No RAG	47.1	82.0	50.4	29.8	54.5
RAG	44.4	76.7	58.0	36.0	55.5
CompAct	38.6	68.9	49.3	30.3	48.4
Astute RAG	52.2	84.1	60.1	44.4	61.7

Table 4: Comparison with evidence filterin on Claude.

Method	NQ	TriviaQA	BioASQ	PopQA	Overall
No RAG	44.8	80.2	45.8	25.3	51.3
RAG	42.7	76.0	55.2	33.7	53.7
CompAct	35.3	65.0	47.6	30.9	46.0
Astute RAG	50.2	81.6	58.0	40.5	59.2

Table 5: Comparison with evidence filterin on Gemini.