

MASS-RAG: Multi-Agent Synthesis Retrieval-Augmented Generation

Xingchen Xiao¹, Heyan Huang^{1*}, Runheng Liu¹, Jincheng Xie²

¹School of Computer Science and Technology, Beijing Institute of Technology

²Department of Mathematical Sciences, Tsinghua University

{xcxiao, hhy63, rhliu}@bit.edu.cn

xiejc22@mails.tsinghua.edu.cn

Abstract

Large language models (LLMs) are widely used in retrieval-augmented generation (RAG) to incorporate external knowledge at inference time. However, when retrieved contexts are noisy, incomplete, or heterogeneous, a single generation process often struggles to reconcile evidence effectively. We propose **MASS-RAG**, a multi-agent synthesis approach to retrieval-augmented generation that structures evidence processing into multiple role-specialized agents. MASS-RAG applies distinct agents for evidence summarization, evidence extraction, and reasoning over retrieved documents, and combines their outputs through a dedicated synthesis stage to produce the final answer. This design exposes multiple intermediate evidence views, allowing the model to compare and integrate complementary information before answer generation. Experiments on four benchmarks show that MASS-RAG consistently improves performance over strong RAG baselines, particularly in settings where relevant evidence is distributed across retrieved contexts.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable scalability and emergent capabilities (Brown et al., 2020; Wei et al., 2022; Touvron et al., 2023; OpenAI et al., 2024). However, they remain susceptible to generating unreliable or hallucinatory outputs due to limited factual grounding and static knowledge. To address these limitations, Retrieval-Augmented Generation (RAG) has emerged as an effective paradigm to enhance factual reliability and broaden the knowledge scope of LLMs (Huang et al., 2023; Xu et al., 2024b; Asai et al., 2024). Updating the parametric knowledge of LLMs through fine-tuning or continual pre-training is both computationally intensive and costly (Meng

*Corresponding author.

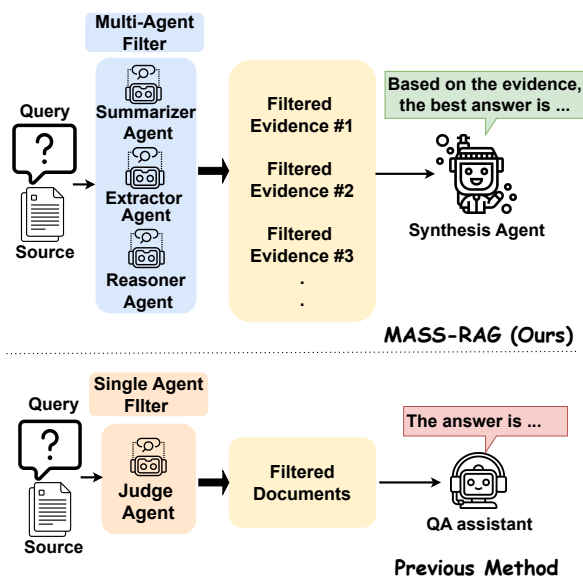


Figure 1: A conceptual illustration of the key idea behind MASS-RAG, contrasting the multi-agent evidence filtering and synthesis with the previous single-filter agent method.

et al., 2023), making RAG integration a practical alternative for maintaining up-to-date knowledge at a lower cost (Balaguer et al., 2024).

Previous efforts on the RAG system have explored either pre-training models from scratch (Borgeaud et al., 2022) or continually pre-training existing models to incorporate external knowledge (Izacard et al., 2022b; Wang et al., 2023a). However, such approaches are computationally expensive and resource-intensive, which limits their scalability in real-world RAG applications (Gao et al., 2024). As a result, recent RAG research has largely focused on combining retrievers for document selection with large language models (LLMs) for generation (Ram et al., 2023; Lin et al., 2023; Shi et al., 2023). By leveraging external non-parametric knowledge sources, these systems enable knowledge updates without retraining the underlying LLMs. Never-

theless, LLMs remain susceptible to irrelevant or redundant retrieved information (Liu et al., 2023; Cho et al., 2023), which can degrade generation quality. Consequently, a key challenge is to ensure that RAG systems produce robust and factually accurate outputs when retrieved contexts are noisy or incomplete due to the inherent nature of the retrieval models (Gao et al., 2024; Asai et al., 2024). While recent multi-agent RAG framework proposed by Chang et al. (2024) introduce agent-based context filtering, they typically rely on a single judge agent operating from a monolithic perspective, which limits the system’s ability to capture complementary or heterogeneous forms of factually relevant evidence.

To address this challenge, we introduce **MASS-RAG**, a Multi-Agent Synthesis framework for retrieval-augmented generation. It operates in a training-free manner and employs multiple agents with specialized roles for context filtering, response generation, and answer synthesis. **MASS-RAG** integrates multiple specialized agents to capture complementary evidence in the retrieved context, thereby improving robustness and factual accuracy through structured multi-agent evidence filtering and synthesis.

Our contributions are as follows:

- **Multi-Agent Evidence Filtering:** We introduce a set of specialized LLM-based filter agents that process retrieved documents from complementary perspectives, distilling denoised and query-relevant evidence in a training-free manner.
- **Multi-Agent Answering and Synthesis:** We propose a synthesis mechanism that integrates structured outputs from multiple filter agents, optionally via intermediate candidate answers, to produce a unified final prediction. This design enables explicit comparison across heterogeneous evidence views and supports task-adaptive configurations.
- **Empirical Validation:** We evaluate MASS-RAG across multiple benchmarks and analysis settings, demonstrating consistent improvements over strong RAG baselines, especially in scenarios requiring the aggregation of complementary evidence.

2 Related Work

Evidence Refinement in RAG A variety of methods have been proposed to optimize the input context for retrieval-augmented generation (RAG). One intuitive line of work focuses on reranking retrieved documents (Glass et al., 2022), which has been widely adopted and extended in subsequent studies (Ram et al., 2023; Asai et al., 2023; Hwang et al., 2024; Xu et al., 2024a). Other approaches aim to refine the input context through query rewriting (Ma et al., 2023; Chan et al., 2024), context filtering (Wang et al., 2023b; Yang et al., 2023), or content compression and summarization (Xu et al., 2023; Hwang et al., 2024). These methods primarily operate by improving the relevance or compactness of the retrieved context prior to generation. Our work follows the line of training-free methods but differs in that we introduce **Multi-Agent Synthesis (MASS)**, **MASS-RAG** adopts a multi-perspective filtering approach, employing multiple specialized agents to distill complementary and denoised evidence from retrieved documents while preserving query-relevant information.

Multi-Agent RAG Recent studies have explored the use of multiple agents to enhance retrieval-augmented generation. Instead of relying on a single retriever-generator pipeline, multi-agent RAG frameworks distribute different sub-processes among specialized agents, such as query reformulation (Chen et al., 2025), query decomposition (Nguyen et al., 2025), and document filtering (Chang et al., 2024). These approaches have demonstrated improved robustness and performance over traditional single-pipeline RAG systems. Our work follows the training-free setting of prior multi-agent RAG methods, but differs in its focus on evidence-centric synthesis. Specifically, **MASS-RAG** introduces a multi-agent synthesis framework in which distinct agents filter retrieved documents from complementary perspectives and generate denoised evidence representations that are subsequently integrated to produce a robust final answer.

3 Multi-Agent Synthesis RAG (MASS-RAG)

3.1 Definition of LLM Agents in MASS-RAG

MASS-RAG decomposes retrieval-augmented generation into three stages: evidence distillation, candidate answer generation, and final answer synthe-

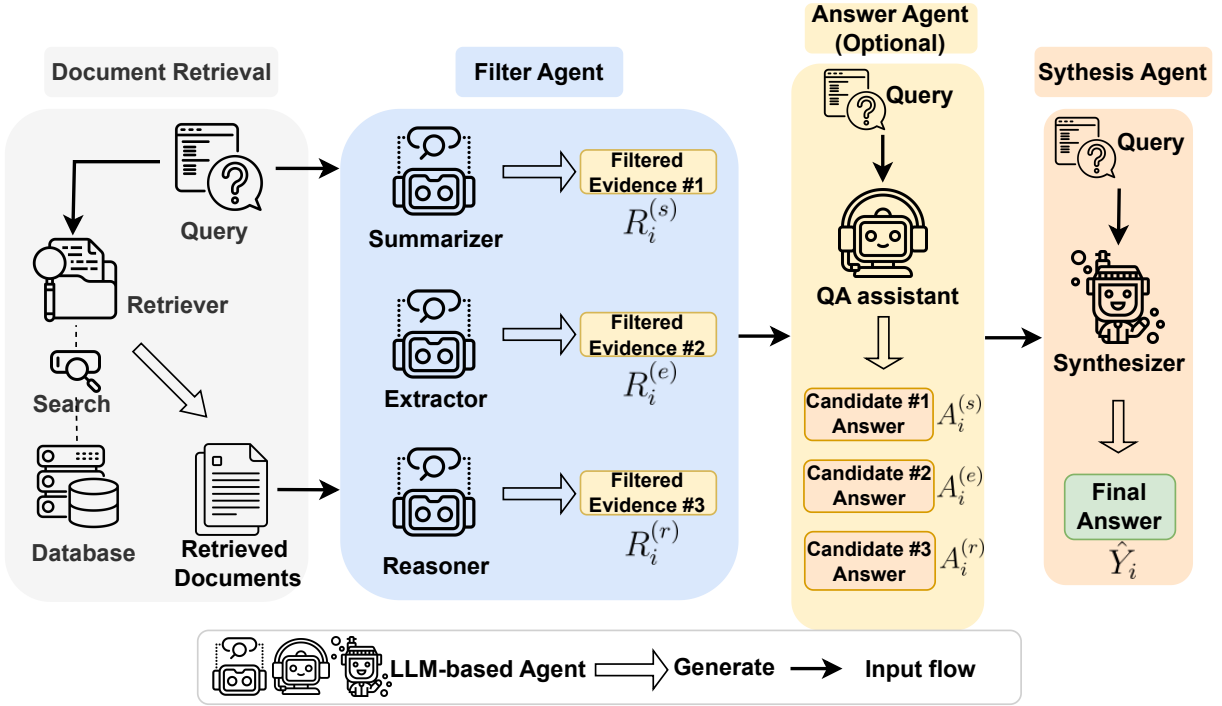


Figure 2: The overall architecture of MASS-RAG, illustrating multi-agent filtering and answer synthesis when the Answer Agent is enabled.

sis. Accordingly, it consists of several specialized LLM agents, including a Summarizer, an Extractor, a Reasoner, a Synthesis Agent, and an optional Answer Agent. Among them, the Summarizer, Extractor, and Reasoner serve as *filter agents* that distill query-relevant evidence from retrieved documents, each emphasizing a distinct perspective on evidence processing.

Summarizer Agent Prior work has shown that compressing retrieved documents can significantly improve RAG performance (Xu et al., 2023; Hwang et al., 2024). Building on this insight, the Summarizer agent condenses the retrieved documents into a concise representation that preserves semantic consistency while highlighting salient, query-relevant information. Formally, the summarization process is defined as:

$$R_i^{(s)} = \mathcal{A}_{\text{sum}}(q_i, D), \quad (1)$$

where $R_i^{(s)}$ denotes the summary-based filtered response.

Extractor Agent The Extractor agent identifies fine-grained factual spans or sentences that explicitly support the answer to q_i without paraphrasing. Unlike the Summarizer, this agent strictly operates in an extractive manner, copying evidence verbatim from the retrieved documents, motivated by

previous work on extractive QA (Chen et al., 2017; Lewis et al., 2021). It is particularly suited for questions whose answers can be directly grounded in explicit textual evidence. This process is expressed as:

$$R_i^{(e)} = \mathcal{A}_{\text{ext}}(q_i, D), \quad (2)$$

where $R_i^{(e)}$ consists of explicit evidence fragments aligned with the query semantics.

Reasoner Agent The Reasoner agent performs reasoning over the retrieved content to infer implicit connections or cross-document evidence that may not be explicitly stated in any single sentence, as motivated by prior work on reasoning and relevance judgment in retrieval-augmented generation (Asai et al., 2023; Chang et al., 2024). The Reasoner Agent leverages the backbone language model to analyze the retrieved documents and explicitly articulate how the evidence supports the query, as well as the inference steps required to arrive at an answer. Formally, it is defined as:

$$R_i^{(r)} = \mathcal{A}_{\text{rea}}(q_i, D), \quad (3)$$

where $R_i^{(r)}$ represents a reasoning-oriented filtered response that connects and interprets evidence across retrieved documents.

Notably, MASS-RAG instantiates multiple agents with explicitly differentiated roles and out-

put constraints. Each agent imposes distinct output constraints that explicitly structure how retrieved documents are processed prior to answer generation. Unlike prior approaches, the Reasoner in MASS-RAG produces an explicit intermediate evidence representation that is decoupled from the final answer.

Answer Agent. Given each filtered response, the Answer Agent independently generates a candidate answer. Specifically, for the outputs from the Summarizer, Extractor, and Reasoner, the Answer Agent computes:

$$A_i^{(j)} = \mathcal{A}_{\text{ans}}(q_i, R_i^{(j)}), \quad j \in \{s, e, r\}, \quad (4)$$

where $A_i^{(s)}$, $A_i^{(e)}$, and $A_i^{(r)}$ denote candidate answers derived from summary-, extraction-, and reasoning-based evidence, respectively.

Generating candidate answers independently encourages the model to commit to a concrete answer under each evidence view. This design is most effective for factoid-style question answering, where candidate answers carry rich semantic content and different evidence views may yield complementary or competing hypotheses. In contrast, for multiple-choice tasks with symbolic option labels (e.g., A/B/C/D), intermediate candidate answers provide limited additional signal. For long-form QA, the candidate-answer step can still be beneficial, though its effectiveness may be constrained by a tendency to focus on a single dominant response. Accordingly, the Answer Agent can be optionally enabled depending on the task characteristics.

Synthesis Agent. The candidate answers $\{A_i^{(s)}, A_i^{(e)}, A_i^{(r)}\}$ may overlap while differing in factual details or reasoning paths. When the Answer Agent is enabled, a **Synthesis Agent** integrates these intermediate answer hypotheses and produces the final output as:

$$\hat{Y}_i = \mathcal{A}_{\text{syn}}(q_i, \{A_i^{(s)}, A_i^{(e)}, A_i^{(r)}\}). \quad (5)$$

When the Answer Agent is disabled, the Synthesis Agent directly operates on the filtered evidence responses generated by the filter agents:

$$\hat{Y}_i = \mathcal{A}_{\text{syn}}(q_i, \{R_i^{(s)}, R_i^{(e)}, R_i^{(r)}\}). \quad (6)$$

The Synthesis Agent is designed to explicitly compare and reconcile structured evidence or answer representations produced by heterogeneous

agents, rather than to simply increase model capacity through additional generation steps. Across both settings, this structured synthesis produces a coherent final answer with improved factual accuracy, completeness, and robustness.

We provide detailed role specifications for all agents in Appendix C.

4 Experiment

We conduct comprehensive experiments and ablation studies to evaluate the performance of MASS-RAG from three perspectives:

- **RQ1:** How does MASS-RAG perform compared to training-based RAG methods and existing training-free RAG approaches, including prior multi-agent systems with single-agent filtering?
- **RQ2:** How do individual filter agents capture factually relevant evidence, and to what extent are their contributions complementary?
- **RQ3:** How does aggregating evidence from multiple filter agents affect the end-to-end performance of the RAG system?

In addition to these research questions, we include targeted ablation studies to examine the robustness of MASS-RAG to retrieval depth and to analyze the role of the optional Answer Agent.

4.1 Datasets

MASS-RAG is evaluated using different models and downstream tasks. We conduct our evaluations under zero-shot setting, where each LLM agent is provided with specific task instructions in the in-context approach (Ram et al., 2023; Chang et al., 2024).

Open-Domain Question Answering Tasks (ODQA). Following previous state-of-the-art (SOTA) work (Chang et al., 2024; Asai et al., 2023), we evaluate MASS-RAG on two ODQA datasets: TriviaQA-unfiltered (Joshi et al., 2017) and PopQA (Mallen et al., 2023). As the testing sets of TriviaQA-unfiltered (open) are not publicly available, to make a fair performance comparison, we use the TriviaQA-unfiltered validation and testing sets provided by existing work (Asai et al., 2023). For PopQA, we use the long-tail subset, consisting of 1,399 rare entity queries whose monthly Wikipedia page views are less than 100.

	TriviaQA	PopQA	ARC-C	ASQA		
	(acc)	(acc)	(acc)	(em)	(rg)	(mau)
<i>Baselines without retrieval</i>						
Llama2 _{7B} *	30.5	14.7	21.8	7.9	15.3	19.0
Llama2 _{1.3B} *	38.5	14.7	29.4	7.2	12.4	16.0
Mistral _{7B} *	54.8	26.2	55.5	11.2	18.1	27.6
Llama3 _{8B} *	68.4	29.2	58.8	19.4	30.3	54.3
<i>Baselines with retrieval (training-based)</i>						
Llama2-FT _{7B} *	57.3	48.7	65.8	31.0	35.8	51.2
Self-RAG _{7B} *	66.4	54.9	67.3	30.0	35.7	74.3
<i>Baselines with retrieval (training-free)</i>						
Llama2 _{7B} *	68.9	50.9	51.0	16.2	23.4	33.1
Llama2 _{1.3B} *	47.0	45.7	26.0	16.3	20.5	24.7
Mistral _{7B} *	69.4	55.5	57.1	32.4	34.8	54.3
Llama3 _{8B} *	73.1	61.8	55.6	37.1	36.5	63.0
Qwen3 _{8B}	72	62.1	<u>86.4</u>	46.1	35.7	16.1
MAIN-RAG-Mistral _{7B} *	71.0	58.9	58.9	35.7	36.2	60.0
MAIN-RAG-Llama3 _{8B} *	<u>74.1</u>	64.0	61.9	39.2	42.0	<u>70.6</u>
MASS-RAG-Llama2_{7B}	68.6	57.8	72.2	36.2	33.3	20.94
MASS-RAG-Mistral_{7B}	73.7	60.9	72.4	41.6	37.1	69.8
MASS-RAG-Llama3_{8B}	76.7	64.2	78.7	<u>47</u>	35.2	55.2
MASS-RAG-Qwen3_{8B}	<u>74.1</u>	63.9	87.3	47.51	<u>37.4</u>	53.5

Table 1: Overall experimental results on four tasks. **Bold numbers** refer to the best performance among baselines without retrieval and training-free baselines, and underline numbers refer to the second-best performance. * indicates concurrent results conducted by recent works or original papers. For the metrics, *acc*, *em*, *rg*, and *mau* denote *accuracy*, *str-em*, *rouge*, and *MAUVE*, respectively.

Long-form Ambiguous Question Answering.

We use a long-form QA dataset with 948 samples, ALCE-ASQA (Stelmakh et al., 2023; Gao et al., 2023), to evaluate MASS-RAG performance in ambiguous questions. Each ambiguous sample in the ASQA dataset may contain multiple pieces of evidence and multiple valid answers, making the task particularly challenging. Following prior work (Asai et al., 2023; Mallen et al., 2023; Chang et al., 2024), we evaluate ASQA using the official metrics, including string exact match (str-em), output diversity measured by MAUVE, and ROUGE.

Closed-set Task We test the performance of our proposed method on the ARC-Challenge dataset (Clark et al., 2018) to assess whether multi-agent context filtering benefits RAG performance on a closed-set reasoning task. We use accuracy as an evaluation metric and report on the test set with 1,172 samples.

4.2 Baselines

Following the previous methods (Asai et al., 2023; Chang et al., 2024), we evaluated the performance of the MASS-RAG and baselines on four benchmarks. The main evaluation metric is accuracy which has been widely used (Asai et al., 2023; Chang et al., 2024).

Baselines without retrievals We compared our method with several publicly available pretrained LLMs, including Llama2_{7B} (Touvron et al., 2023), Llama3_{8B} (Grattafiori et al., 2024), and Mistral_{7B} (Jiang et al., 2023).

Baselines with retrievals We compared our method with both training-based method and training-free method, which incorporate retrieval either during inference or through the training process. For training-based baselines, we consider Self-RAG (Asai et al., 2023) and the Llama2-FT_{7B} which is the Llama2_{7B} fine-tuned on the same

dataset used by Self-RAG but without the reflection tokens or retrieved passages. For training-free baselines, we used the models we included in the baselines without retrievals and we also introduce the latest model Qwen3_{8B} as reference (Yang et al., 2025). Eventually, we compared our method with the previous SOTA (state-of-the-art) multi-agent RAG method, the MAIN-RAG proposed by Chang et al. (2024). We also compared our method with standard RAG baselines which do not require additional training and generally prepend the top retrieved documents to the query as the pre-trained LLM input, using the same retriever as all other baselines we include.

4.3 Main Results (RQ1)

Table 1 reports the accuracy results of all methods. MASS-RAG consistently outperforms both training-based and training-free baselines across all evaluated tasks. To ensure a fair comparison with Self-RAG, which is trained on Llama2_{7B}, we evaluate MASS-RAG using the same backbone. Under this setting, MASS-RAG achieves absolute accuracy improvements of up to 3.3%, 5.3%, 7.3%, and 20.7% on the four selected tasks, respectively.

We further compare MASS-RAG with MAIN-RAG using Llama3_{8B} as the backbone. As shown in Table 1, MASS-RAG yields improvements of up to 3.5%, 0.3%, 27.1%, and 19.9% across the same tasks, demonstrating the effectiveness of the proposed multi-agent synthesis strategy. We visualize these comparisons in Appendix B.

4.4 Experimental Settings

MASS-RAG is compatible with different pre-trained LLMs as backbone models and operates in a fully training-free manner. We use the retrieved documents released by Asai et al. (2023) for all experiments. For each query, up to 20 documents are retrieved and ranked using the pre-trained Contriever trained on MS MARCO (Izacard et al., 2022a). This setting is also adopted by Self-RAG and MAIN-RAG, enabling a fair and controlled comparison. All models are decoded using greedy generation, with temperature set to 0 and *top-p* set to 1.0, ensuring deterministic and reproducible outputs. The number of top retrieved documents is a hyperparameter in our experiments. We use the top-5 retrieved documents for Llama2_{7B} and the top-10 retrieved documents for other models, following the settings adopted by the corresponding

Method	TQA (acc)	PopQA (acc)	ARC- C (acc)	ASQA (em)
<i>Retrieved Docs = 5</i>				
Self-RAG	66.4	54.9	67.3	30.0
MASS-RAG(Llama2 _{7B})	68.6	57.8	72.2	36.2
MASS-RAG(Llama3 _{8B})	75.8	61.5	79.9	45.2
MASS-RAG(Mistral _{7B})	72.7	55.7	72.3	39.5
<i>Retrieved Docs = 10</i>				
MAIN-RAG(Llama3 _{8B})	74.1	64.0	61.9	39.2
MASS-RAG(Llama3 _{8B})	76.7	64.2	78.7	47.0
MASS-RAG(Mistral _{7B})	73.7	60.9	72.4	41.6

Table 2: Performance of MASS-RAG on four downstream tasks with different numbers of retrieved documents.

baselines.

For TriviaQA, PopQA, and ASQA, we instantiate all five agents following the standard MASS-RAG pipeline shown in Figure 2 to ensure a consistent experimental setup across QA benchmarks. For ARC-Challenge, we instantiate only four agents and exclude the optional Answer Agent. This setting adapts to the multiple-choice nature of ARC-Challenge, where answers are restricted to symbolic option labels (e.g., A, B, C, D) and thus lack semantic granularity. The structure of ARC-Challenge does not naturally support intermediate candidate-answer generation. Accordingly, for ARC-Challenge, the final prediction is directly synthesized by the Synthesis Agent from the filtered responses of the three filter agents.

4.5 Ablation Study and Analysis

Ablation Study on # of Retrieved Documents

Due to limitations in retrieval ranking (Asai et al., 2024), factually relevant evidence may not always appear among the top-ranked documents, making RAG performance sensitive to the number of retrieved documents. To assess the robustness to retrieval depth, we evaluate MASS-RAG under varying numbers of retrieved documents, as shown in Table 2.

Across the evaluated tasks, MASS-RAG exhibits stable performance trends as the retrieval depth varies, demonstrating robustness to changes in the number of retrieved documents. Notably, when fewer documents are selected as retrieval results, MASS-RAG instantiated with Llama3_{8B} consistently outperforms MAIN-RAG on TriviaQA, ARC-Challenge, and ASQA, indicat-

Dataset	Model	w/ Answer	w/o Answer
TriviaQA	Mistral	77.5	76.0
	Llama3	79.9	79.3
PopQA	Mistral	60.9	61.9
	Llama3	64.2	62.6
ASQA	Mistral	41.6	43.9
	Llama3	47.0	48.0

Table 3: Effect of the Answer Agent across different backbone models. We report performance with and without the Answer Agent on short-form QA (TriviaQA, PopQA) and long-form QA (ASQA).

ing that **MASS-RAG** remains effective even with limited retrieved context, where robust evidence filtering and synthesis play a more critical role.

Ablation Study on the effect of Answer Agent

We analyze the role of the Answer Agent by ablating this component across different backbone models and datasets. In MASS-RAG, three filter agents (Summarizer, Extractor, and Reasoner) are first applied to distill complementary evidence views, and the Answer Agent then independently generates one candidate answer for each filtered response, resulting in three candidate answers corresponding to the three filter agents.

As shown in Table 3, incorporating the Answer Agent consistently improves performance on factoid QA benchmarks, including TriviaQA and PopQA, across both Mistral and Llama3 backbones. This suggests that independently generating answer hypotheses from multiple evidence views helps reduce ambiguity and enables effective comparison during the synthesis stage.

In contrast, the Answer Agent yields marginal or no improvement on ASQA. Unlike factoid QA, ASQA requires aggregating a diverse set of factual statements into a comprehensive long-form answer. In this setting, early commitment to individual candidate answers can reduce the flexibility of integrating information distributed across multiple evidence views. Consistent with this observation, we disable the Answer Agent in our main experiments on ARC-Challenge, where answers are constrained to symbolic options and do not benefit from intermediate answer hypothesis generation. Overall, these results highlight the task-dependent utility of the Answer Agent and demonstrate the flexibility of MASS-RAG, where this component can be optionally enabled depending on task characteristics.

Construction of the Uniquely Attributable Subset (RQ2)

To better quantify and analyze the contributions of individual filter agent responses, we process the TriviaQA and PopQA datasets using the outputs of MASS-RAG. Specifically, we identify a subset of questions whose ground-truth evidence is captured by exactly one of the three filter agents, meaning that each such question is captured by a single type of filter agent. This allows us to quantify how many questions can be answered exclusively based on a particular filter agent response. Through this analysis, we demonstrate that different filter agents provide complementary factual evidence that is uniquely captured by each agent, and that all three agents are indispensable, as each contributes distinct and non-overlapping factual evidence to the overall MASS-RAG system.

For PopQA, following the setup in Section 4.1, we start from 1,399 questions and construct a **Uniquely Attributable Subset** by retaining only questions whose ground-truth evidence is captured by exactly one filter agent, while excluding questions captured by multiple agents or by none. This subset enables focused analysis of questions uniquely handled by each agent. The same procedure is applied to TriviaQA for consistency. This analysis reveals the distinct contribution of each filter agent and highlights the complementarity among their responses. Accordingly, the size of the **Uniquely Attributable Subset** varies with the backbone model, resulting in 609 (TriviaQA) and 108 (PopQA) samples for Llama3_{8B}, and 559 (TriviaQA) and 85 (PopQA) samples for Mistral_{7B}.

Effectiveness of Individual Filter Agents (RQ2)

To evaluate the effectiveness of each type of filter agent response ($R_i^{(s)}$, $R_i^{(e)}$, $R_i^{(r)}$) introduced in Section 3.1, we define **Evidence Coverage Rate (ECR)**, inspired by the Answer Inclusion Rate (AIR) proposed by Ruan et al. (2024), as the proportion of questions for which a given filter agent response contains ground-truth evidence. It is defined as:

$$\text{ECR}(A_F) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[A_F(q) \text{ relevant}] \quad (7)$$

In equation 7, A_F denotes a single filter agent or a set of filter agents whose responses are jointly considered, and $\mathbf{1}[\cdot]$ is an indicator function that equals 1 if $A_F(q)$ contains ground-truth evidence for question q , and 0 otherwise.

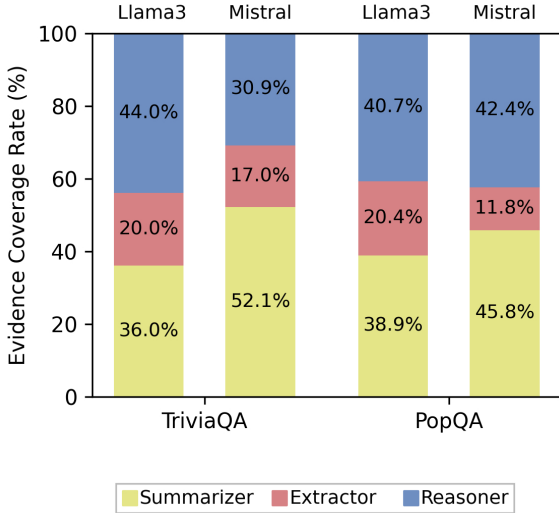


Figure 3: Evidence Coverage Rate (ECR) of each filter response on the **Uniquely Attributable Subset** of TriviaQA and PopQA using Llama3_{8B} and Mistral_{7B} with the top-10 retrieved documents.

As shown in Figure 3, different filter agents capture distinct and complementary subsets of factually relevant evidence. Different agents tend to excel on different subsets of questions, supporting the need for multi-agent filtering, since a single agent alone is insufficient to reliably cover all relevant evidence. Additional case studies are provided to qualitatively demonstrate the complementarity among filter agents (see Appendix D).

Impact of Multi-Agent Evidence Aggregation (RQ3) Building on the evidence-level robustness observed in RQ2, we investigate whether aggregating evidence from multiple agents enhances the end-to-end performance of the RAG system under noisy retrieval.

As described in Section 3.1, MASS-RAG employs a synthesis agent \mathcal{A}_{syn} to produce the final answer by aggregating candidate answers generated from the responses of the filter agents ($R_i^{(s)}$, $R_i^{(e)}$, $R_i^{(r)}$). We evaluate this process using answer accuracy on TriviaQA and PopQA. As shown in Tables 4 and 5, the synthesis agent does not always attain the highest evidence coverage, reflecting its distinct role. The summarization agent achieves high coverage, fulfilling its design objective; however, its output serves as an intermediate representation that may include multiple pieces of relevant evidence without resolving their correctness or mutual consistency. Notably, the intermediate responses produced by the filter agents often achieve

Agent	TriviaQA (acc)	PopQA (acc)
Summarizer	73.7	65.7
Extractor	68.1	61.5
Reasoner	72.8	63.0
Synthesis	76.7	64.2

Table 4: Accuracy (acc) of each filter response on TriviaQA and PopQA using Llama3_{8B} with the top-10 retrieved documents.

Agent	TriviaQA (acc)	PopQA (acc)
Summarizer	73.8	61.2
Extractor	66.9	57.8
Reasoner	71.9	62.4
Synthesis	73.6	60.9

Table 5: Accuracy (acc) of each filter response on TriviaQA and PopQA using Mistral_{7B} with the top-10 retrieved documents.

higher accuracy, indicating that these agents are effective at filtering retrieved documents and exposing answerable content supported by the evidence. In this sense, the filtered responses can be viewed as an approximate upper bound on what the model could potentially answer given the retrieved context. The role of the Synthesis Agent, however, is to produce a single unified answer rather than retain multiple partially correct or ambiguous responses. Whether this upper bound can be reached in the final prediction therefore depends on the base model’s ability to consolidate and resolve the filtered evidence. This observation is consistent with our ablation results, where stronger backbone models are better able to approach this upper bound than weaker ones (e.g. Llama3 vs Mistral).

5 Conclusion

We propose MASS-RAG, a multi-agent synthesis framework for retrieval-augmented generation. Experimental results indicate that role-specialized multi-agent filtering captures complementary and factually relevant evidence from retrieved contexts, and that synthesizing these filtered representations leads to more reliable final predictions. Overall, our results suggest that explicitly structuring evidence processing into multi-agent filtering followed by synthesis is a practical design choice for improving

the RAG systems.

Limitations

Our study is designed to isolate and analyze the effects of multi-agent evidence filtering and synthesis within a retrieval-augmented generation framework. To enable controlled comparisons, we adopt fixed retrieval pipelines and pretrained backbone language models throughout our experiments. Accordingly, components such as retriever design, reranking strategies, and decoding-level variations are treated as fixed rather than independent variables in this work.

MASS-RAG is further evaluated under standard inference-time configurations without task-specific parameter adaptation. This design choice allows us to focus on the interaction between agent-level evidence processing and synthesis, while leaving open how learning-based or adaptive mechanisms could complement the proposed framework in future studies.

The modular design of MASS-RAG also makes it amenable to future extensions along orthogonal dimensions, such as retrieval or learning-based adaptation, without altering its core synthesis mechanism.

Ethical Statement

This work studies a training-free, multi-agent framework for retrieval-augmented generation, focusing on how retrieved evidence is filtered and synthesized by large language models. The proposed method does not introduce new data sources, supervision signals, or learning objectives beyond those used in existing RAG systems, and therefore inherits the ethical considerations associated with the underlying pretrained models and retrieval corpora.

As with other retrieval-based generation approaches, MASS-RAG may surface inaccuracies or biases present in retrieved documents or in the backbone language models. While the proposed multi-agent filtering and synthesis mechanism aims to improve factual grounding and robustness under noisy retrieval, it does not guarantee the correctness or neutrality of generated outputs. Accordingly, the system should not be used as a sole source of truth in high-stakes or safety-critical applications without appropriate human oversight.

We conduct all experiments using publicly available benchmarks and released retrieval results, and

do not involve personal data, user profiling, or deployment-facing evaluation. Our goal is to provide a controlled analysis of multi-agent evidence processing rather than to advocate immediate real-world deployment.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). *Preprint*, arXiv:2310.11511.
- Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. 2024. [Reliable, adaptable, and attributable language models with retrieval](#). *Preprint*, arXiv:2403.03187.
- Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. 2024. [Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture](#). *Preprint*, arXiv:2401.08406.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. [Improving language models by retrieving from trillions of tokens](#). *Preprint*, arXiv:2112.04426.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- 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](#). *Preprint*, arXiv:2404.00610.
- Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2024. [Main-rag: Multi-agent filtering retrieval-augmented generation](#). *Preprint*, arXiv:2501.00332.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. [Reading wikipedia to answer open-domain questions](#). *Preprint*, arXiv:1704.00051.

- Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. 2025. [Improving retrieval-augmented generation through multi-agent reinforcement learning](#). *Preprint*, arXiv:2501.15228.
- Sukmin Cho, Jeongyeon Seo, Soyeong Jeong, and Jong C. Park. 2023. [Improving zero-shot reader by reducing distractions from irrelevant documents in open-domain question answering](#). *Preprint*, arXiv:2310.17490.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *Preprint*, arXiv:1803.05457.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). *Preprint*, arXiv:2305.14627.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#). *Preprint*, arXiv:2312.10997.
- Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. [Re2g: Retrieve, rerank, generate](#). *Preprint*, arXiv:2207.06300.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *Preprint*, arXiv:2311.05232.
- Taeho Hwang, Soyeong Jeong, Sukmin Cho, SeungYoon Han, and Jong C. Park. 2024. [Dslr: Document refinement with sentence-level re-ranking and reconstruction to enhance retrieval-augmented generation](#). *Preprint*, arXiv:2407.03627.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. [Unsupervised dense information retrieval with contrastive learning](#). *Preprint*, arXiv:2112.09118.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. [Atlas: Few-shot learning with retrieval augmented language models](#). *Preprint*, arXiv:2208.03299.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). *Preprint*, arXiv:1705.03551.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K ttler, Mike Lewis, Wen tau Yih, Tim Rockt schel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Preprint*, arXiv:2005.11401.
- Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2023. [Radtit: Retrieval-augmented dual instruction tuning](#). *Preprint*, arXiv:2310.01352.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. [Lost in the middle: How language models use long contexts](#). *Preprint*, arXiv:2307.03172.
- Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. [Query rewriting for retrieval-augmented large language models](#). *Preprint*, arXiv:2305.14283.
- Alex Mullen, 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](#). *Preprint*, arXiv:2212.10511.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. [Locating and editing factual associations in gpt](#). *Preprint*, arXiv:2202.05262.
- Thang Nguyen, Peter Chin, and Yu-Wing Tai. 2025. [Maraag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning](#). *Preprint*, arXiv:2505.20096.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. [In-context retrieval-augmented language models](#). *Preprint*, arXiv:2302.00083.

- Qian Ruan, Ilya Kuznetsov, and Iryna Gurevych. 2024. [Are large language models good classifiers? a study on edit intent classification in scientific document revisions](#). *Preprint*, arXiv:2410.02028.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. [Replug: Retrieval-augmented black-box language models](#). *Preprint*, arXiv:2301.12652.
- Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2023. [Asqa: Factoid questions meet long-form answers](#). *Preprint*, arXiv:2204.06092.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. 2023a. [Instructretro: Instruction tuning post retrieval-augmented pretraining](#). *Preprint*, arXiv:2310.07713.
- Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. [Learning to filter context for retrieval-augmented generation](#). *Preprint*, arXiv:2311.08377.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models](#). *Preprint*, arXiv:2206.07682.
- Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. [Recomp: Improving retrieval-augmented lms with compression and selective augmentation](#). *Preprint*, arXiv:2310.04408.
- Zhipeng Xu, Zhenghao Liu, Yibin Liu, Chenyan Xiong, Yukun Yan, Shuo Wang, Shi Yu, Zhiyuan Liu, and Ge Yu. 2024a. [Activerag: Revealing the treasures of knowledge via active learning](#). *Preprint*, arXiv:2402.13547.
- Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. [Hallucination is inevitable: An innate limitation of large language models](#). *Preprint*, arXiv:2401.11817.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. [Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter](#). *Preprint*, arXiv:2310.18347.

Appendix

A Additional Experimental Setup

Table 6 conclude the computational setup used for all experiments. Models with 7B or 8B parameters can be executed on a single 24GB GPU, while models larger than 13B parameters typically require at least two 24GB GPUs. To ensure a fair comparison with prior work and to facilitate economical reproducibility, all experiments are conducted using bfloat16 precision, consistent with the settings adopted by *Self-RAG* (Asai et al., 2023). The total runtime of each experiment may vary depending on system-level factors such as GPU power limits, runtime load, and driver versions.

Device Attribute	Specification
Computing Infrastructure	GPU
GPU Model	NVIDIA RTX 4090
Number of GPUs	4
GPU Memory	24 GB

Table 6: Computing Device for the experiments.

B MASS-RAG Performance Comparison with *SELF-RAG* and *MAIN-RAG*

Figure 4 presents a visualized comparison between **MASS-RAG** and two representative state-of-the-art baselines, *Self-RAG* and *MAIN-RAG*, across four benchmark datasets. In the left panel, **MASS-RAG** is compared with *Self-RAG* under a shared Llama2_{7B} backbone, while the right panel reports comparisons with *MAIN-RAG* using Llama3_{8B} to ensure backbone consistency. Across all benchmarks, **MASS-RAG** consistently outperforms both baselines, with particularly notable gains on reasoning-intensive datasets such as ARC-C and ASQA. The annotated improvements highlight that **MASS-RAG** yields stable performance gains under controlled backbone settings, suggesting that the observed improvements are attributable to the proposed multi-agent synthesis framework rather than differences in model variation.

C Agent Role Specifications

Each agent in **MASS-RAG** is instantiated with a lightweight, role-specific instruction that constrains its output objective and format (e.g., extractive, abstractive, or reasoning-oriented), thereby inducing distinct intermediate representations from the same retrieved context.

Summarizer Agent The Summarizer is responsible for condensing the retrieved documents into a concise, abstractive summary that highlights query-relevant information. Its output preserves the semantic content of the retrieved context while removing irrelevant or redundant details, and does not attempt to directly answer the query. Even when the retrieved documents do not contain a complete answer, the Summarizer still produces a partial summary of any potentially relevant information.

Extractor Agent The Extractor identifies and selects fine-grained textual spans from the retrieved documents that explicitly support answering the query. It operates in a strictly extractive manner, copying evidence verbatim from the context without paraphrasing or reinterpretation. When multiple relevant spans exist, the Extractor may return several evidence fragments, reflecting all directly grounded information available in the retrieved documents.

Reasoner Agent The Reasoner examines the retrieved documents to identify implicit or cross-document connections that are not explicitly stated in any single sentence. Rather than producing a final answer, it articulates how different pieces of retrieved evidence could be combined or interpreted to support answering the query, including the assumptions or inference steps involved. This agent focuses on reasoning over the retrieved context without introducing external knowledge.

Answer Agent (optional) The Answer Agent generates a concise candidate answer based on a single filtered evidence representation produced by one filter agent. Its role is to translate the provided evidence into a concrete answer hypothesis, without introducing information beyond the given evidence. When the evidence is partial or incomplete, the Answer Agent produces a best-effort answer that reflects the available information rather than refusing to respond. This agent does not perform cross-evidence aggregation and is applied indepen-

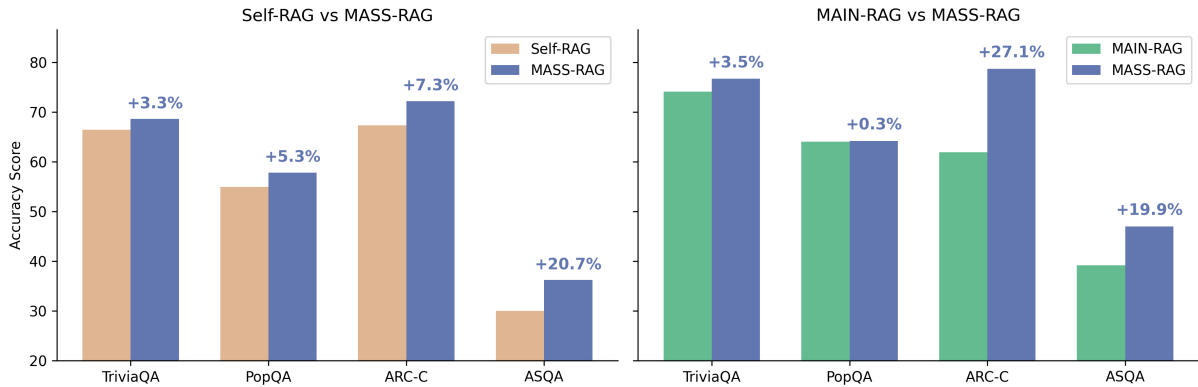


Figure 4: Performance comparison of MASS-RAG with state-of-the-art baselines across four benchmark datasets. The left panel reports results compared with *Self-RAG* using Llama2_{7B}, while the right panel shows comparisons with *MAIN-RAG* based on Llama3_{8B}. In both cases, the backbone models are kept consistent to ensure a fair comparison.

dently to each filter agent output, serving as an intermediate hypothesis generator rather than the final decision module.

Synthesis Agent The Synthesis Agent produces the final output by consolidating structured inputs from multiple agents. Depending on the task setting, these inputs may consist of filtered evidence representations or intermediate candidate answers. The Synthesis Agent performs comparative reasoning across its inputs, integrates complementary information, resolves inconsistencies, and generates a coherent final answer without referring to individual agents explicitly.

These role specifications are intentionally lightweight and transferable across backbone language models.

D Case Study Of MASS-RAG

We provide additional case studies comprising six examples from Figure 5 to 10, including three cases evaluated with Mistral_{7B} on PopQA and three cases evaluated with Llama3_{8B} on TriviaQA. Each case illustrates a distinct scenario in which only one or two filter agents are able to capture the relevant factual or implicit evidence. Implicit evidence refers to evidence that does not explicitly contain the gold answer, but nonetheless allows the model to infer the correct answer without requiring additional information. All cases are analyzed used on top-10 retrieved documents setting and the gold answers for each sample are underlined.

E Runtime Cost Analysis Of MASS-RAG

In order to analyze the total runtime cost of MASS-RAG, we assume that each agent run costs 100 tokens. We use x as the default run time for one agent call. In table 7, we compare each sample run under different settings.

Setting	Run Cost
MASS-RAG w/ Answer Agent	$8x$
MASS-RAG w/o Answer Agent	$4x$

Table 7: Runtime Cost Analysis for different settings.

Case Study 1 Question: Who was the composer of Overture in G major?
<p>[Summarizer Agent] (Evidence Include): Anton Bruckner composed the Overture in G minor, not G major. The Overture in G major was composed by <u>Luigi Cherubini</u>.</p> <p>[Extractor Agent] (No Evidence): Anton Bruckner, specifically the Overture in G minor, WAB 98. However, the question asked for Overture in G major, and there is no evidence in the provided context that Bruckner composed an Overture in G major.</p> <p>[Reasoner Agent] (No Evidence): The Context provides information about several composers and their works, but none of them are identified as the composer of the Overture in G major. However, it does mention Anton Bruckner and his Overture in G minor, WAB 98, which was composed during the same time period as the Overture in G major. Inference steps required: 1. Recognize that the Context provides information about various composers and their works, but not the composer of the Overture in G major. 2. Understand that the Overture in G minor by Bruckner was composed during the same time period as the Overture in G major.</p>
<p>[Answer Agent] (Candidate Answers):</p> <p>1: Luigi Cherubini was the composer of the Overture in G major. (Correct Answer)</p> <p>2: Based on the provided evidence, there is no information that Anton Bruckner composed an Overture in G major. Therefore, I cannot confirm the composer of this specific work. (NOT Answer)</p> <p>3: The Context does not provide the identity of the composer of the Overture in G major. However, it's worth noting that Anton Bruckner composed the Overture in G minor during the same time period. This information might suggest a possible connection or similarity in style between the two composers, but it does not definitively answer who the composer of the Overture in G major is. (NOT Answer)</p>
<p>[Synthesis Agent] (Final Answer): <u>Luigi Cherubini</u> was the composer of the Overture in G major. (Correct Answer)</p>

Figure 5: MASS-RAG Case Study 1 - Dataset: PopQA, Model: Mistral_{7B}

Case Study 2 Question: In what country is Ciepień?
<p>[Summarizer Agent] (No Evidence): Ciepień is twinned with a location, but the context does not specify the country where Ciepień is located.</p> <p>[Extractor Agent] (Evidence Include): Ciepień is twinned with Ciechanów, which is located in <u>Poland</u>. However, the context does not provide specific information about where Ciepień is located within <u>Poland</u>.</p> <p>[Reasoner Agent] (No Evidence): The Context provides information about various locations, but it does not directly mention "Ciepień" in relation to a country. However, the context does mention "Ciechanów" and its twin towns, which could potentially provide a clue. Inference steps required:</p> <ol style="list-style-type: none"> 1. Identify the locations that Ciechanów is twinned with (Ciepień and Gmina Zbójno). 2. Investigate whether any of these locations are countries. 3. If Ciepień is not mentioned as a location, check if it could be a part of one of the twin towns (Gmina Zbójno).
<p>[Answer Agent] (Candidate Answers):</p> <p>1: The provided context does not specify the country where Ciepień is located. Therefore, it cannot be definitively stated. (Not Answer)</p> <p>2: Ciepień is located in <u>Poland</u>. (Correct Answer)</p> <p>3: Based on the provided context, it is not possible to definitively determine the country of Ciepień. Ciechanów is a town in Poland, and it is twinned with Ciepień and Gmina Zbójno. However, neither Ciepień nor Gmina Zbójno are mentioned as countries in the provided context. Therefore, further research would be needed to determine the country of Ciepień. (Incomplete Answer)</p>
<p>[Synthesis Agent] (Final Answer): Ciepień is located in <u>Poland</u>, as suggested by Candidate Answer 2. However, Candidate Answer 3 provides additional context that Ciepień is twinned with Ciechanów, a town in Poland. Since Ciechanów is in Poland, it is reasonable to infer that Ciepień is also likely in Poland. Therefore, the final answer is: Ciepień is located in <u>Poland</u>. (Correct Answer)</p>

Figure 6: MASS-RAG Case Study 2 - Dataset: PopQA, Model: Mistral_{7B}

Case Study 3 Question: What genre is Drive On?
<p>[Summarizer Agent] (No Evidence): Drive (2011 film) is a neo-noir, art house feature.</p> <p>[Extractor Agent] (No Evidence): Drive (2011 film): Drive has been described as a tough, hard-edged, neo-noir, art house feature, extremely violent and very stylish, with European art and grindhouse influences.</p> <p>[Reasoner Agent] (Evidence Include): The Context provides information about the Drive (2011 film) and its soundtrack, but there's no direct mention of the genre of the Drive (2011 film) or the Drive On (song). However, the film Drive is described as a neo-noir, art house feature, extremely violent and very stylish, with European art and grindhouse influences. The soundtrack, which includes <u>electronic-pop</u> songs, is also described as integral to the film's experience.</p>
<p>[Answer Agent] (Candidate Answers):</p> <p>1: The genre of Drive (as a film) is neo-noir and art house. (Incorrect Answer)</p> <p>2: Drive is a neo-noir film. (Incorrect Answer)</p> <p>3: The provided context does not directly mention the genre of the song "Drive On." However, given that the film Drive (2011) has a soundtrack with electronic-pop songs and the film itself has neo-noir, art house, European art, and grindhouse influences, it's possible that the genre of "Drive On" could be related to these genres, particularly <u>electronic pop</u> or <u>synth-pop</u>, due to its inclusion in the film's soundtrack. However, without specific information about "Drive On," this is an inference. (Correct Answer)</p>
<p>[Synthesis Agent] (Final Answer): Drive is a neo-noir film, so it's possible that the genre of "Drive On" (if it's a song from the film's soundtrack) could be related to neo-noir, <u>electronic-pop</u>, or <u>synth-pop</u>. However, without specific information about "Drive On," this is an inference. (Correct Answer)</p>

Figure 7: MASS-RAG Case Study 3 - Dataset: PopQA, Model: Mistral_{7B}

Case Study 4 Question: What relation is Joey Richardson to Vanessa Redgrave?
<p>[Summarizer Agent] (Evidence Include): Joely Richardson is Vanessa Redgrave's <u>daughter</u>.</p> <p>[Extractor Agent] (Implicit Evidence): Vanessa Redgrave is the mother of Joely Richardson.</p> <p>[Reasoner Agent] (No Evidence): To determine the relation between Joey Richardson and Vanessa Redgrave, we need to examine the context and look for any information that might connect the two individuals. The context provides information about the Redgrave family, including their relationships and descendants. We can see that Vanessa Redgrave is mentioned as the sister of Lynn Redgrave and the mother of Natasha Richardson and Joely Richardson. However, there is no mention of a person named Joey Richardson in the context. Therefore, we can conclude that there is no information in the context that directly connects Joey Richardson to Vanessa Redgrave. However, we can make an inference that Joey Richardson is likely a relative of Vanessa Redgrave, given the context's information about the Redgrave family and their relationships.</p>
<p>[Answer Agent] (Candidate Answers):</p> <p>1: Vanessa Redgrave's <u>daughter</u>. (Correct Answer)</p> <p>2: Joely Richardson is the <u>daughter</u> of Vanessa Redgrave. (Correct Answer)</p> <p>3: Joey Richardson is likely a relative of Vanessa Redgrave, possibly a child or descendant of Natasha Richardson or Joely Richardson. However, the exact relation is not specified in the context.. (NOT Answer)</p>
<p>[Synthesis Agent] (Final Answer): Vanessa Redgrave's <u>daughter</u>. (Joely Richardson) is the correct answer. Vanessa Redgrave's <u>daughter</u> is Joely Richardson. (Correct Answer)</p>

Figure 8: MASS-RAG Case Study 4 - Dataset: TriviaQA, Model: Llama3_{8B}

Case Study 5 Question: The US No 1 single from Diana Ross's platinum album "Diana" is "Endless Love", a duet with Lionel Richie.

[Summarizer Agent] (No Evidence):

The US No 1 single from Diana Ross's platinum album "Diana" is "Endless Love", a duet with Lionel Richie.

[Extractor Agent] (Evidence Include):

Evidence: "Upside Down" eventually was chosen by the label and; "Endless Love", composed for Brooke Shields' film of the same name, by and performed with Commodores front man, Lionel Richie. Ross and Richie performed the song on the telecast for the 54th Academy Awards, where the song was nominated in the category of "Best Song". The success of the single launched Richie's successful solo career. Ross has also ventured into acting, with a Golden Globe Award and Academy Award-nominated performance for her performance in the film "Lady; "I'm Still Waiting" and "Chain Reaction," which were both solo #1 hits for Diana Ross in the UK. International UK Diana Ross & the Supremes:

[Reasoner Agent] (Implicit Evidence):

To answer the question, we need to examine the context and look for information related to Diana Ross's platinum album "Diana" and a US No 1 single. We can find this information in the section about the album "Diana". The album "Diana" was released in May 1980 and was a platinum album. The inference step would be to look for a US No 1 single from this album. The context provides sufficient evidence for this step, as it mentions that the album "Diana" was a platinum album and that it introduced Ross to a new generation of fans worldwide. However, it does not explicitly mention a US No 1 single from this album. We would need to make an assumption that the album "

[Answer Agent] (Candidate Answers):

1: "Endless Love", a duet with Lionel Richie. (Incorrect Answer)

2: The US No 1 single from Diana Ross's platinum album Diana was "Upside Down". It was also a No 1 hit in the UK. The song was chosen by the label and was composed by and performed with Lionel Richie. The success of the single launched Richie's successful solo career. Ross has also ventured into acting, with a Golden Globe Award and Academy Award-nominated performance for her performance in the film "Lady Sings the Blues". "I'm Still Waiting" and "Chain Reaction," which were both solo #1 hits for Diana Ross in the UK. International UK Diana Ross & the Supremes: The No 1 single from Diana Ross's platinum album Diana was "Upside Down". (Correct Answer)

3: The US No 1 single from Diana Ross's platinum album "Diana" is "Upside Down". It was released in 1980 and became a huge hit, reaching number one on the US Billboard Hot 100 chart. The song was a single from the album "Diana", which was a platinum album and introduced Ross to a new generation of fans worldwide. "Upside Down" is considered one of Ross's signature songs and a classic of the disco era. It has been covered and sampled by many artists over the years and remains a beloved and iconic song in popular culture. The song's success can be attributed to its catchy melody, memorable lyrics, and Ross's powerful vocals. (Correct Answer)

[Synthesis Agent] (Final Answer):

The US No 1 single from Diana Ross's platinum album Diana was "Upside Down". It was released in 1980 and became a huge hit, reaching number one on the US Billboard Hot 100 chart. The song was a single from the album "Diana", which was a platinum album and introduced Ross to a new generation of fans worldwide. "Upside Down" is considered one of Ross's signature songs and a classic of the disco era. (Correct Answer)

Figure 9: MASS-RAG Case Study 5 - Dataset: TriviaQA, Model: Llama3_{8B}

Case Study 6 What was the first British football team to win the European Cup?
<p>[Summarizer Agent] (No Evidence): Manchester United became the first English club to win the European Cup in 1968, beating Benfica 4–1 in the final.</p> <p>[Extractor Agent] (No Evidence): Evidence: Manchester United became the first English club to win the European Cup in 1968, beating Benfica 4–1 in the final.</p> <p>[Reasoner Agent] (Evidence Include): To answer the question, we need to identify the first British football team to win the European Cup. The relevant information in the context is the sentence: "<u>Celtic</u> became the first British club to win the Cup in 1967, beating Italian side Inter who had won 2 of the previous 3 finals 2–1 in Lisbon." This sentence provides the necessary information to answer the question. However, it does not mention the first English team to win the European Cup. We need to look for another sentence that mentions the first English team to win the European Cup.</p>
<p>[Answer Agent] (Candidate Answers):</p> <p>1: Manchester United became the first English club to win the European Cup in 1968, beating Benfica 4–1 in the final. Therefore, Manchester United was the first British football team to win the European Cup. (Incorrect Answer)</p> <p>2: Manchester United became the first English club to win the European Cup in 1968, beating Benfica 4–1 in the final. Therefore, Manchester United was the first British football team to win the European Cup. (Incorrect Answer)</p> <p>3: <u>Celtic</u> became the first British club to win the European Cup in 1967, beating Italian side Inter 2–1 in Lisbon. However, the first English team to win the European Cup was Manchester United in 1968, when they defeated Benfica 4–1 in the final. (Correct Answer)</p>
<p>[Synthesis Agent] (Final Answer): <u>Celtic</u> became the first British club to win the European Cup in 1967, beating Italian side Inter 2–1 in Lisbon. However, the first English team to win the European Cup was Manchester United in 1968, when they defeated Benfica 4–1 in the final. Manchester United was the first English team to win the European Cup, but <u>Celtic</u> was the first British team. Therefore, the first British football team to win the European Cup was <u>Celtic</u>. (Correct Answer)</p>

Figure 10: MASS-RAG Case Study 6 - Dataset: TriviaQA, Model: Llama3_{8B}