Sentence Filtering for Adaptive Retrieval-Augmented Generation in Multi-Hop Question Answering

Omala Koralage
School of Computing Technologies
RMIT University
Naarm / Melbourne, Australia
omala.vishmini.koralage@student.rmit.edu.au

Damiano Spina
School of Computing Technologies
RMIT University
Naarm / Melbourne, Australia
damiano.spina@rmit.edu.au

Abstract

Adaptive Retrieval-Augmented Generation (RAG) systems such as Self-aware Knowledge Retrieval (SeaKR) typically retrieve passages iteratively based on the Large Language Model (LLM)'s internal uncertainty, but use the full retrieved passages as external context without filtering irrelevant content which can reduce answer generation quality. In this work, we examine how sentence-level filtering methods applied to retrieved passages affect answer generation in uncertainty-based adaptive RAG for multi-hop question answering. We compare two filtering approaches: within-passage sentence filtering (WPSF), which filters sentences within each retrieved passage independently and passes all filtered passages to the LLM, and cross-passage sentence filtering (CPSF), which selects the most relevant sentences globally across all retrieved passages and passes a single fused passage to the LLM. Preliminary results on the HotpotQA and 2WikiMultiHopQA multi-hop question answering datasets show that WPSF is significantly more effective than CPSF, yet neither differs significantly from the unfiltered baseline (SeaKR). A qualitative analysis indicates that the results depend largely on how filtering is integrated into SeaKR: the filtered text is not kept for later reasoning steps, which leads to repeated searches and early stops, and CPSF, by selecting a fixed number of sentences across all passages, can leave out the sentence needed for the answer.

Keywords: cross-passage filtering, within-passage filtering, uncertainty-based RAG, multi-hop question answering

1 Introduction

Large language models (LLMs) can generate answers that have no factual basis, a behavior known as hallucination (Klesel and Wittmann, 2025). Retrieval-augmented generation (RAG) reduces this risk by retrieving documents or passages from an external knowledge source and providing them to the model when it generates an answer (Lewis et al., 2020b; Klesel and Wittmann, 2025). Complex questions such as multi-hop questions remain challenging for RAG, because when answering them the model must retrieve and reason over several pieces of supporting evidence rather than a single passage (Tang and Yang, 2024; Zhuang et al., 2024).

Multi-hop questions cannot usually be answered by retrieving once with the question alone. What the model needs to look up often depends on what it has already worked out (Trivedi et al., 2023). For this reason, iterative and adaptive RAG methods retrieve several times while the model reasons, collecting evidence step by step (Trivedi et al., 2023; Jiang et al., 2023; Su et al., 2024b). These methods mainly differ in how they decide when to retrieve. Adaptive-RAG (Jeong et al., 2024) chooses a strategy based on how complex the question is, and Amber (Qin et al., 2025) uses the knowledge gathered in earlier steps. SeaKR (Yao et al., 2025b) takes a different approach and retrieves only when the model itself is uncertain, which it measures from the model's internal states.

In most of these systems, the retrieved passages are simply added to the model's input (Ram et al., 2023), and irrelevant or repeated text in them can mislead the model (Hwang et al., 2024; Yao et al., 2025a). To address this, several studies filter retrieved content before it reaches the model. Some filter whole passages, for example by having several LLM agents score each passage (Chang et al., 2025) or by estimating how much useful information it contains (Yan et al., 2026). Others filter at the sentence level, splitting each passage into sentences and keeping only those relevant to the question (Hwang et al., 2024; Yao et al., 2025a), and some combine document- and sentence-level filtering (Masoumi et al., 2026; Choi et al., 2024). These methods are designed for static RAG, where passages are retrieved once for the question. Only

a few studies filter content in multi-step RAG, for example by topic (Rezaei et al., 2024) or with a fine-tuned LLM (Qin et al., 2025), and uncertainty-based adaptive RAG methods such as SeaKR still give the model each retrieved passage in full.

In this work, we apply sentence-level filtering to passages retrieved adaptively for multi-hop questions. We integrate two filtering methods into SeaKR. Within-passage sentence filtering (WPSF) selects the relevant sentences within each retrieved passage, and cross-passage sentence filtering (CPSF) selects the most relevant sentences across all passages retrieved at a given step. Comparing the two shows whether sentences are better selected within each passage or across passages. We evaluate both methods on HotpotQA and 2WikiMultiHopQA. Neither method differs significantly from the unfiltered baseline, but WPSF is significantly more effective than CPSF, suggesting that selecting sentences within each passage is preferable.

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) was introduced by Lewis et al. (2020b) as a method to extend large language models with external knowledge at inference time, improving factual accuracy, and reducing hallucinations by retrieving relevant passages and prepend them on the model input. This retrieve-then-read pipeline demonstrated strong performance on open-domain QA tasks without requiring the model to memorize all knowledge in its parameters. Ram et al. (Ram et al., 2023) further showed that retrieval can be performed at fixed token intervals during generation without modifying the LM architecture, demonstrating that even rule-based periodic retrieval using standard retrievers improves the language model perplexity and QA accuracy across model sizes. However, these approaches retrieve a fixed number of passages regardless of whether retrieval is actually needed, and a single or fixed retrieval step is insufficient for tasks that require multiple reasoning steps over different pieces of evidence.

2.2 Adaptive and Iterative Retrieval

To address the limitations of fixed retrieval, dynamic and iterative retrieval methods have been proposed that decide when and what to retrieve based on the generation context. IRCoT (Trivedi et al., 2023) interleaves retrieval with chain-of-thought reasoning, using each generated reasoning sentence as a retrieval query and accumulating retrieved passages to guide subsequent steps, substantially improving multi-hop QA over one-step retrieve-and-read baselines. FLARE (Jiang et al., 2023) takes a confidence-based approach, monitoring token-level probabilities during generation and triggering retrieval only when the model produces low-confidence tokens, avoiding unnecessary retrieval when the model is already certain. DRAGIN (Su et al., 2024b) extends this by considering the LLM's information needs across the entire generation context rather than only the most recent sentence, identifying retrieval triggers and query terms from the full context to more precisely address what the model actually needs. DRAD (Su et al., 2024a) focuses on entity-level hallucination, detecting when named entity probabilities or entropy indicate potential factual errors and retrieving targeted context at those points to correct generation before it continues. SELF-RAG (Asai et al., 2024) takes a training-based approach, fine-tuning a single LM to adaptively retrieve passages on demand and reflect on both retrieved passages and its own outputs using special reflection tokens, enabling controllable inference without a fixed retrieval schedule. RA-ISF (Liu et al., 2024) further combines iterative task decomposition with self-feedback, breaking complex questions into sub-tasks and refining retrieval across iterations to reduce the impact of irrelevant retrieved content on model performance.

2.3 Document Filtering and Refinement in RAG

A key challenge in RAG systems is that retrieved documents often contain irrelevant or noisy content that degrades generation quality. Several approaches have been proposed to address this through document filtering and refinement at different levels of granularity.

Hwang et al. (2024) propose DSLR, which decomposes retrieved passages into individual sentences, scores and filters out irrelevant ones, and reconstructs them back into coherent passages. However, this approach is applied only within a standard RAG setting. Similarly, Yao et al. (2025a) propose ParetoRAG, which applies the Pareto principle to identify and retain the most informative sentences from retrieved passages while discarding less relevant ones. This reduces token usage by up to 60% while maintaining answer quality. However, like

DSLR, ParetoRAG is designed for standard single-step RAG and does not address the challenges of multi-hop reasoning. Multi-granular approaches span both standard and adaptive settings. Masoumi et al. (2026) propose MG-CRAG, which extends corrective RAG by applying two sequential evaluators: a passage-level scorer that discards globally irrelevant documents, followed by a sentence-level scorer that removes noisy sentences within retained passages. This hierarchical refinement is applied within a standard RAG setting and does not address multi-hop reasoning. Choi et al. (2024) propose MGFiD, which integrates multi-granular filtering into the Fusion-in-Decoder architecture through multi-task learning jointly training a passage reranker and a sentence classifier, then aggregating retained evidence sentences into an anchor vector that guides decoding. This filtering is applied in standard open-domain QA and does not extend to adaptive or multi-hop settings.

Table 1 compares these approaches with the methods studied in this work. Adaptive retrieval methods do not filter the retrieved content, and filtering methods are applied only in single-step retrieval. WPSF and CPSF combine both.

3 Approach

3.1 Proposed Dynamic RAG Architecture

Yao et al. (2025b) propose SeaKR (Self-aware Knowledge Retrieval), which extracts self-aware uncertainty from the internal states of the LLM specifically the Gram determinant of intermediate layer representations to determine when retrieval is necessary. At each reasoning step, if the model's uncertainty exceeds a threshold, SeaKR triggers retrieval and re-ranks retrieved passages using the same uncertainty signal to select the most informative passage, avoiding unnecessary retrievals when the model is already confident. SeaKR is evaluated on both complex multi-hop QA datasets (2WikiMultiHopQA, HotpotQA, IIRC) and simple QA datasets (NQ, TriviaQA, SQuAD), outperforming existing adaptive RAG baselines. We adopt SeaKR as our baseline as it provides a principled, training-free adaptive retrieval mechanism grounded in the model's internal uncertainty, making it a strong and reproducible foundation for evaluating the effect of document filtering independently of retrieval decisions.

3.2 Within-Passage Sentence Filtering (WPSF)

We apply two sentence-level filtering approaches to the passages adaptively retrieved by SeaKR. In the within-passage filtering approach, each retrieved passage is split into individual sentences, and each sentence is scored by computing the cosine similarity between its embedding and the embedding of the current question. Sentences whose similarity score falls below a threshold T are discarded, and the remaining sentences are reconstructed into a filtered passage that is passed to the LLM for reasoning. This ensures that only sentences directly relevant to the current question are retained within each retrieved passage, reducing noise in the generation context without modifying the retrieval process itself.

3.3 Cross-Passage Sentence Filtering (CPSF)

The second approach, cross-passage fusion, extends filtering beyond individual passages by operating across all passages retrieved at a given reasoning step. All sentences from all retrieved passages are pooled together, and each sentence is scored by computing the cosine similarity between its embedding and the embedding of the current question. The top-K highest-scoring sentences are selected from this pooled set and concatenated into a single fused context that is passed to the LLM for reasoning.

4 Experimental Setup

We now describe our experimental setup, including the datasets and evaluation metrics (Section 4.1) and the implementation details (Section 4.2).

4.1 Datasets and Metrics

Datasets. We evaluate on two multi-hop question answering development datasets.[1] **HotpotQA** (Yang et al., 2018) consists of 7,405 questions that require reasoning over multiple Wikipedia paragraphs, including bridge and comparison question types. **2WikiMultiHopQA** (Ho et al., 2020) consists of 12,576 questions constructed to require reasoning across two Wikipedia articles, covering five question types designed to probe multi-hop reasoning capabilities. For both datasets, we use the few-shot prompts in SeaKR's implementation, which follow Trivedi et al. (2023): each example

[1]For this is ongoing work, we opted to use the dev sets in both datasets, and left the test splits for the final experiments.

Approach	Post-Retrieval Filtering	Multi-Step Retrieval	Adaptive Retrieval	Uncertainty-Based
IRCoT (Trivedi et al., 2023)		✓		
FLARE (Jiang et al., 2023)		✓	✓	✓
DRAGIN (Su et al., 2024b)		✓	✓	✓
SeaKR (Yao et al., 2025b)		✓	✓	✓
DSLR (Hwang et al., 2024)	✓			
ParetoRAG (Yao et al., 2025a)	✓			
MG-CRAG (Masoumi et al., 2026)	✓			
MGFiD (Choi et al., 2024)	✓			
WPSF (ours)	✓	✓	✓	✓
CPSF (ours)	✓	✓	✓	✓

Table 1: Comparison of adaptive retrieval and post-retrieval filtering approaches. WPSF and CPSF apply post-retrieval filtering within uncertainty-based adaptive retrieval.

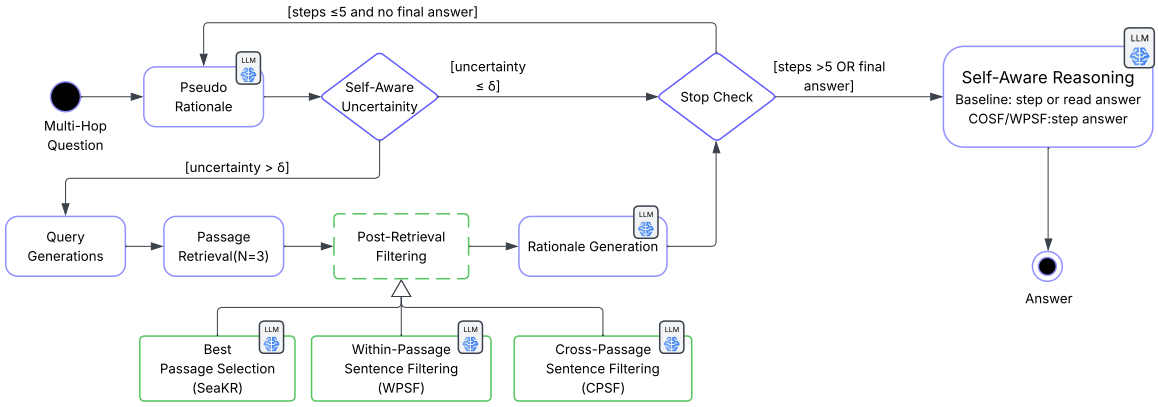


Figure 1: Overview of the dynamic RAG architecture.

question is answered with step-by-step reasoning ending in "So the answer is", and the LLM follows this format to generate its reasoning steps and final answer. The prompt template and example demonstrations are given in Appendix B.

Evaluation Metrics. We evaluate usingYao et al. (2025b)'s implementation of Exact Match (EM) and token-level F1 score. EM measures whether the predicted answer string exactly matches the gold answer, and F1 measures token-level overlap between the predicted and gold answer strings.

Statistical Tests. We used Student's paired t-test with significance level $\alpha = 0.05$ and Bonferroni correction for multiple comparisons.

4.2 Implementation Details

We used Yao et al. (2025b)'s implementation of SeaKR[2] with LLaMA-2-Chat with 7 billion parameters as the backbone LLM, Elasticsearch's implementation of Okapi BM25 (Robertson et al., 1994) as the retrieval model, and the English Wikipedia dump from December 2018 as the external knowledge source, as in the original SeaKR configuration. The number of retrieved passages per step is set to $N = 3$, the self-aware uncertainty threshold is set to $\delta = -6.0$, and the eigen-score uncertainty estimation is adopted as-is from the original SeaKR paper. Yao et al. (2025b) run SeaKR with up to seven reasoning steps and five retrieved documents per question in their released code, whereas we use up to five reasoning steps and three retrieved documents. The model, retriever, N, δ and these limits are the same for all three conditions.

All three conditions follow the pipeline in Figure 1. At each reasoning step, the LLM first generates the next reasoning step without retrieval, a *pseudo rationale*. If its self-aware uncertainty is at most δ, it is kept as the next rationale. Otherwise, a search query is generated from the pseudo rationale, N passages are retrieved, the passages are processed by the condition-specific method described below, and the LLM generates the rationale from the processed text. As in SeaKR, reasoning stops when

[2] https://github.com/THU-KEG/SeaKR

the LLM states the final answer ("So the answer is") or when the maximum number of reasoning steps is reached. The final answer is then produced by self-aware reasoning, which differs between the conditions.

SeaKR (Baseline). The baseline uses SeaKR's self-aware re-ranking (Yao et al., 2025b): the LLM generates a rationale from each of the N retrieved passages separately, and the passage that leaves the model least uncertain is kept together with its rationale. Kept passages are added to the context of later steps, so evidence accumulates as reasoning proceeds, and retrieval stops once three passages have been kept. In self-aware reasoning, the final answer is produced in two ways: a step answer, generated from the question and the rationales, and a read answer, generated from the question and all kept passages. The answer with the lower uncertainty is returned.

WPSF. All N retrieved passages are retained. Each passage is split into sentences, and each sentence is scored by the cosine similarity between the BART embeddings (Lewis et al., 2020a) of the sentence and the question. Sentences scoring below $T = 0.3$ are discarded, and the remaining sentences of all passages are passed to the LLM together, which generates a single rationale from them. The threshold was set without tuning. The filtered text is used only at the step at which it was retrieved and is not kept for later steps. In self-aware reasoning, only the step answer is generated, and it is returned as the final answer.

CPSF. The sentences of all N retrieved passages are pooled into a single set. Each sentence is scored in the same way as in WPSF, and the $K = 5$ highest-scoring sentences are joined into a single context, from which the LLM generates a single rationale. The value of K was set without tuning. As in WPSF, the filtered text is not kept for later steps, and only the step answer is generated and returned as the final answer.

5 Results

Table 2 reports EM and F1 scores for all three runs on both datasets. WPSF obtains the highest scores on every metric and dataset. Its advantage over CPSF is statistically significant on HotpotQA for both EM and F1, and on 2WikiMultiHopQA for F1. This suggests that selecting sentences within each passage is more effective than pooling sentences across passages. Compared with the SeaKR baseline (EM = 0.2799, F1 = 0.3347 on 2WikiMultiHopQA), WPSF scores are higher and CPSF scores are lower on both datasets, but these differences are not statistically significant.

6 Qualitative Analysis

We conduct a qualitative analysis on the 2WikiMultiHopQA results. Questions are grouped into five outcome categories based on which run produced the correct answer:

- **SeaKR only correct:** answered correctly by SeaKR but incorrectly by both WPSF and CPSF.
- **WPSF only correct:** answered correctly by WPSF but incorrectly by both SeaKR and CPSF.
- **CPSF only correct:** answered correctly by CPSF but incorrectly by both SeaKR and WPSF.
- **WPSF and CPSF correct, SeaKR wrong:** answered correctly by both WPSF and CPSF but incorrectly by SeaKR.
- **All Conditions Incorrect:** answered incorrectly by all three runs.

For each category, we examine five questions from the 2WikiMultiHopQA dataset, using the reasoning logs of each run. The questions and the answers of each run are listed in Appendix C.

6.1 Cases Where Only SeaKR Is Correct

Table 3 lists the questions in this group: four bridge comparisons (Questions 1, 2, 3 and 5), which need four facts, the person linked to each film and a date for each person, and one comparison (Question 4), which needs only the dates of two named people. In this group, the errors of WPSF and CPSF came mainly from how queries are formed and from not keeping retrieved text between steps, not from the filtering itself. Because the query is built from the confident words of the pseudo rationale, an uncertain name is dropped and the same query can be issued again; without the earlier retrieved text, the model could not change its next step, and reasoning stopped early. Filtering caused errors only in CPSF, whose fixed number of sentences left out the needed sentence. Several errors were shared by all

Run	HotpotQA		2WikiMultiHopQA	
	EM	F1	EM	F1
SeaKR (Yao et al., 2025b)	0.2849	0.3987	0.2799	0.3347
WPSF	**0.2902***	**0.4031***	**0.2840**	**0.3417***
CPSF	0.2783	0.3917	0.2784	0.3321

Table 2: Effectiveness in terms of EM and F1 on the HotpotQA and 2WikiMultiHopQA dev sets. Bold indicates the highest score per metric. * indicates a statistically significant difference from CPSF.

systems: the search returned films or people with similar names, and cut-off names produced weak queries. These errors all occurred on the bridge comparisons, while gathering the four facts. On the comparison, all three systems reasoned correctly, and WPSF and CPSF failed only when the final answer was generated again from the rationales. SeaKR was not better informed in these questions; it was correct because it keeps its selected passages and generates a second answer from them, which in some questions recovered the correct answer.

6.2 Cases Where Only WPSF Is Correct

Table 4 lists the questions in this group: three bridge comparisons (Questions 6, 7 and 8) and two compositional questions (Questions 9 and 10), which need a person and then a place linked to that person. On the bridge comparisons, no system found all four facts, and WPSF's correct answers mostly did not follow from its reasoning: in Questions 6 and 7, its rationales pointed to the other film, and the final answer, generated again from the rationales, flipped to the correct one. The same step flipped CPSF's reasoning to the wrong film in Question 8. The other errors on these questions were shared: the search returned a different person with the same name or a film with a similar title, and dates were compared the wrong way round. On the compositional questions, the difference came from how retrieved text reaches the model. In Question 10, one retrieved passage named a different composer; SeaKR, which judges passages one at a time, and CPSF, whose top five sentences included that passage, both followed it, while WPSF, which gives the model all filtered passages together, kept the correct composer. In Question 9, SeaKR's correct first guess was uncertain, and its query, built from the confident words only, was reduced to "buried" and then drifted to the wrong place. SeaKR's choice between its two final answers returned the wrong one in Questions 6 and 8, the opposite of what it did in Section 6.1.

6.3 Cases Where Only CPSF Is Correct

Table 5 lists the questions in this group: three bridge comparisons (Questions 11, 12 and 13), which need the director of each film and a death date for each director, and two compositional questions (Questions 14 and 15), which need a person and then a country or place linked to that person. The main issue in this group is the early stop seen in Section 6.1, which ended WPSF's reasoning in four of the five questions. In Questions 11, 12 and 13, WPSF wrote a rationale about the other film; its next pseudo rationale guessed a new director for the first film, but because the query keeps only confident words, the guessed name was dropped, the query repeated, and reasoning stopped. On the bridge comparisons, CPSF's correct answers did not come from better evidence: it never had both directors and both death dates, so they were correct by chance. On the compositional questions, two patterns appeared. In Question 14, SeaKR and WPSF first guessed "United Kingdom" correctly, but the guess was uncertain, so they retrieved a passage stating that the director was born in London, England, and answered "England", which exact-match evaluation counts as wrong. In Question 15, WPSF stated the husband and his burial place in one rationale and never searched for the place, while CPSF stated only the husband and then searched for his burial place, which is the only case in this group where CPSF reasoned correctly. SeaKR's choice between its two final answers again returned the wrong one in Question 13.

6.4 Cases Where WPSF and CPSF Are Correct and SeaKR Is Not

Table 6 lists the questions in this group: four bridge comparisons (Questions 16, 17, 18 and 19), which need the director of each film and a death date for each director, and one comparison (Question 20), which needs only the death dates of two named

people. On the bridge comparisons, the main issue is retrieval, not filtering: five of the eight correct directors were not retrieved by any system. The query keeps only the confident words of the pseudo rationale, which can leave it too general; in Question 17, the query for If I Were Free was reduced to "Free directed" and returned a passage about football. Uncertainty estimation also played a part: when the model was confident in a wrong director, as WPSF and CPSF were in Question 19, it did not search at all. WPSF and CPSF therefore reached the correct film from wrong directors, either because two errors cancelled out, such as wrong dates compared the wrong way round, or because the final answer named a film without a comparison after CPSF repeated its query and stopped in Question 16; these answers were correct by chance. SeaKR reached its limit of three kept passages in four of the five questions, so it could not gather all four facts, and in Questions 17 and 19 its choice between its two final answers returned the wrong one. On the comparison, SeaKR retrieved both dates and reasoned correctly, but its step answer, generated again from the rationales, named the wrong person and was returned, the same issue as for WPSF and CPSF in Question 4.

6.5 Cases Where All Conditions Are Incorrect

Table 7 lists the questions in this group: four compositional questions (Questions 21, 22, 23 and 24), which need a person and then a date or place linked to that person, and one bridge comparison (Question 25). The errors are mostly shared by the three systems and are not caused by filtering. The main issue on the compositional questions is the second fact. The passages retrieved for the person did not contain the date or place, and the model either wrote the person and the fact in the same rationale or was confident enough not to search, so the fact was never retrieved and came from the model itself. In Question 21, the next pseudo rationale was empty, the query fell back to the question, repeated, and reasoning stopped in all three systems, including SeaKR. Query formation also caused errors: in Question 24, the query lost the film's year and retrieved a passage about a fictional centaur, and in Question 25, it was reduced to "born 1970", after which all three systems compared the dates the wrong way round. In Question 22, WPSF and CPSF identified the correct director while SeaKR did not: its query contained its own wrong guess, and choosing one passage at a time kept a passage supporting that guess. The one error specific to filtering is in Question 21, where CPSF's five sentences included a repeated sentence and a sentence about another person's father, and left out the sentence naming the correct father.

6.6 Discussion

Our aim is to understand how sentence filtering behaves when applied inside SeaKR. We therefore keep SeaKR unchanged and focus on the issues that come from the filtering methods and from how they are integrated into SeaKR. For each finding, we describe the change that informs it.

Finding 1: CPSF's fixed number of sentences can leave out the needed sentence. In Questions 2, 5 and 21, the sentence needed for the answer was retrieved but not among CPSF's five, and in Question 10 a misleading sentence was among them. WPSF, which has no fixed limit, kept the needed sentence in the same questions, which is consistent with WPSF being significantly more effective than CPSF. *Next step:*we will tune the number of sentences CPSF keeps on a held-out development set, instead of fixing it at five, and also test choosing it per question.

Finding 2: Repeated sentences take CPSF's places. Passages in the corpus overlap, so the same sentence can appear in several retrieved passages. In Section 6.1, only 20 of the 25 sentences selected by CPSF were distinct, and in Questions 5 and 21 a repeated sentence took a place while the needed sentence was left out. *Next step:* we will remove repeated sentences before CPSF selects its top sentences.

Finding 3: WPSF's threshold removes little text. At $T = 0.3$, WPSF kept 99 of the 110 sentences we checked, so its input stays close to the unfiltered passages, which may explain why it does not differ significantly from SeaKR. *Next step:* we will tune T in a waiting set and report how much text each setting removes.

Finding 4: Filtered text is not kept between steps. In our integration, WPSF and CPSF use the filtered text only at the step where it was retrieved, whereas SeaKR keeps its selected passages for later steps. When a rationale does not resolve the model's uncertainty, the next pseudo rationale and its query repeat, and reasoning stops early. This happened in Questions 1, 2, 11, 12, 13, 15 and 16. *Next step:* we

will keep the filtered text from each step for later steps, in the same way SeaKR keeps its passages.

Finding 5: The filtering methods produce only one final answer. SeaKR generates a second answer from its kept passages and returns the one with lower uncertainty, which recovered the correct answer in Questions 2 and 5. Our integration uses only the answer generated from the rationales, which in Question 4 contradicted the correct reasoning. *Next step:* once the filtered text is kept between steps, we will also generate SeaKR's second answer from it, so that the filtering methods use SeaKR's full answer selection.

Limitations and Future work

This study has several limitations. First, all experiments use a single backbone LLM and a single retriever, so it is unclear whether the results generalize to LLMs of other sizes and architectures, or to other retrievers.

Second, the number of sentences kept by CPSF ($k = 5$) and the WPSF threshold ($T = 0.3$) were fixed and not tuned. Our qualitative analysis shows that, with a fixed number of sentences, CPSF can leave out the sentence it needs, and that the same sentence can be selected more than once because passages in the corpus overlap. WPSF at $T = 0.3$ kept most sentences, so its input stayed close to the unfiltered passages.

Third, we limited each question to five reasoning steps and three retrieved documents, whereas the released SeaKR code uses seven reasoning steps and five retrieved documents. Many 2WikiMultiHopQA questions compare two entities and need four facts, which may not fit within five reasoning steps. In addition, the baseline reached the three-document limit in 30.9% of 2WikiMultiHopQA questions, compared with 11.0% of HotpotQA questions. These lower limits may explain why our 2WikiMultiHopQA baseline scores are lower than those reported by Yao et al. (2025b).

Fourth, and most importantly, WPSF and CPSF differ from the baseline in more than the filtering step. The baseline keeps its selected passages for later reasoning steps, whereas WPSF and CPSF use their filtered text only at the step where it was retrieved; as a result, the model can repeat the same search and stop reasoning early with facts missing. WPSF and CPSF also produce a single final answer, whereas the baseline produces two and returns the one the model is more confident about. Therefore, differences in results between the baseline and the filtering methods may be caused by these other differences, not only by filtering.

In future work, we plan to evaluate the filtering methods with other LLMs and retrievers, and to tune k and T on a held-out development set before reporting results on a separate test set. We will also use the limits of the released SeaKR code (seven reasoning steps and five retrieved documents), so that questions needing four facts have enough steps to be answered and our baseline matches the original setting. In addition, we will keep the filtered text for later reasoning steps and generate SeaKR's second answer from it, so that filtering is the only difference between the baseline and the filtering methods and its effect can be measured directly. The other changes informed by our qualitative analysis are described in Section 6.6.

7 Conclusion

In this paper, we investigated post-retrieval sentence filtering as a strategy for improving multi-hop question answering in adaptive RAG. We proposed two filtering methods, WPSF and CPSF, and added them to SeaKR, an adaptive RAG system. We then compared them with SeaKR without filtering on two benchmark datasets, HotpotQA and 2WikiMultiHopQA. WPSF scored slightly higher than SeaKR and CPSF slightly lower on both datasets, but neither differed significantly from SeaKR; only the difference between WPSF and CPSF was significant. Our qualitative analysis showed that these results depend less on filtering itself than on how it was integrated into SeaKR: the filtered text was not kept for later reasoning steps, which led to repeated searches and early stops, and CPSF's fixed number of sentences sometimes left out the sentence needed for the answer. Our next step is therefore to keep the filtered text across reasoning steps, so that the effect of sentence filtering in adaptive RAG can be measured on its own.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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. 2025. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622, Vienna, Austria. Association for Computational Linguistics.

Eunseong Choi, Hyeri Lee, and Jongwuk Lee. 2024. Multi-granularity guided fusion-in-decoder. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2201–2212, Mexico City, Mexico. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Taeho Hwang, Soyeong Jeong, Sukmin Cho, SeungYoon Han, and Jong Park. 2024. DSLR: Document refinement with sentence-level re-ranking and reconstruction to enhance retrieval-augmented generation. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 73–92, Bangkok, Thailand. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong 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 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Zhengbao Jiang, Frank 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, Singapore. Association for Computational Linguistics.

Michael Klesel and H. Felix Wittmann. 2025. Retrieval-augmented generation (RAG). *Business & Information Systems Engineering*, 67(4):551–561.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. RA-ISF: Learning to answer and understand from retrieval augmentation via iterative self-feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4730–4749, Bangkok, Thailand. Association for Computational Linguistics.

Negin Masoumi, Omid Davar, and Mahdi Eftekhari. 2026. Mg-crag: fusion of multi-granular retrieval evaluators in corrective rag with weakly supervised fine-tuning. *Knowledge and Information Systems*, 68(1):149.

Qitao Qin, Yucong Luo, Yihang Lu, Zhibo Chu, Xiaoman Liu, and Xianwei Meng. 2025. Towards adaptive memory-based optimization for enhanced retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7991–8004, Vienna, Austria. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Mohammad Reza Rezaei, Maziar Hafezi, Amit Satpathy, Lovell Hodge, and Ebrahim Pourjafari. 2024. AT-RAG: An adaptive rag model enhancing query efficiency with topic filtering and iterative reasoning. *Preprint*, arXiv:2410.12886.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024a. Mitigating entity-level hallucination in large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, pages 23–31, Tokyo, Japan. Association for Computing Machinery.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024b. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013.

Yixuan Tang and Yi Yang. 2024. Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling (COLM'24)*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Shuhuan Yan, Wei Wang, Hongwei Chen, and Jie Wang. 2026. IKF-RAG: Intrinsic knowledge-aware and learning-based filtering for enhancing retrieval-augmented generation. *The Journal of Supercomputing*, 82(4):200.

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

Ruobing Yao, Yifei Zhang, Shuang Song, Yuhan Liu, Neng Gao, and Chenyang Tu. 2025a. ParetoRAG: Leveraging sentence-context attention for robust and efficient retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4137–4151, Suzhou, China. Association for Computational Linguistics.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, and Juanzi Li. 2025b. SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27022–27043, Vienna, Austria. Association for Computational Linguistics.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. EfficientRAG: Efficient retriever for multi-hop question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3411, Miami, Florida, USA. Association for Computational Linguistics.

A Declaration of Use of Generative AI

The authors used generative AI tools to assist with writing and editing this paper. All experimental design, analysis, results, and conclusions are the authors' own.

B Prompt Template

All three conditions use the same few-shot prompt from SeaKR's implementation, which follows Trivedi et al. (2023). The prompt contains ten example questions with step-by-step answers, the retrieved text (Context), an instruction, the question, and the reasoning steps generated so far. Figure 2 shows the WPSF prompt for a 2WikiMultiHopQA question at a step where retrieval was triggered. In WPSF, the Context contains the filtered sentences of each of the $N = 3$ retrieved passages, given together in one prompt. The other conditions differ only in the Context: in CPSF, it contains a single passage made of the top-K sentences, and in the baseline, it contains one retrieved passage at a time, together with the passages kept from earlier steps. Without retrieval, the Context is left out.

Question: Who was born first out of Martin Hodge and Ivania Martinich?

Answer: Martin Hodge was born on 4 February 1959. Ivania Martinich was born on 25 July 1995. Thus, 4 February 1959 is earlier than 25 July 1995 and Martin Hodge was born first. So, the answer is Martin Hodge.

Question: When did the director of film Hypocrite (Film) die?

Answer: The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So, the answer is 19 June 2013.

Question: Are both Kurram Garhi and Trojkrsti located in the same country?

Answer: Kurram Garhi is in the country of Pakistan. Trojkrsti is in the country of Republic of Macedonia. Thus, they are not in the same country. So, the answer is no.

Question: Do director of film Coolie No. 1 (1995 Film) and director of film The Sensational Trial have the same nationality?

Answer: Coolie No. 1 (1995 film) was directed by David Dhawan. The Sensational Trial was directed by Karl Freund. David Dhawan's nationality is India. Karl Freund's nationality is Germany. Thus, they do not have the same nationality. So, the answer is no.

Question: Who is Boraqchin (Wife Of Ögedei)'s father-in-law?

Answer: Boraqchin is married to Ögedei Khan. Ögedei Khan's father is Genghis Khan. Thus, Boraqchin's father-in-law is Genghis Khan. So, the answer is Genghis Khan.

... (5 more sample shots in the same format) ...

Context:

[1] House of the Black Death is a 1965 horror film directed by Harold Daniels.
[2] Harold Daniels was an American film director.
[3] Harold Daniels was born in 1903. Answer in the same format as before.

Question: Which film has the director born later, Tell It To The Bees or House Of The Black Death?

Answer: Tell It To The Bees was directed by Annabel Jankel.

Figure 2: Example WPSF prompt for a 2WikiMultiHopQA question (five of the ten example questions shown). The Context holds the filtered sentences of the three retrieved passages; these sentences are illustrative.

C Questions Used in the Qualitative Analysis (2WikiMultiHopQA dev set)

No.	Question	SeaKR ✓	WPSF ✗	CPSF ✗	Gold Answer
1	Which film has the director born later, Tell It To The Bees or House Of The Black Death?	Tell It To The Bees	House Of The Black Death	House Of The Black Death	Tell It To The Bees
2	Which film has the director who died later, Swarg Narak or Antardhan?	Swarg Narak	Antardhan	Antardhan	Swarg Narak
3	Which film has the director born earlier, Chaitrada Chandrama or Amelia Lópes O'Neill?	Amelia Lópes O'Neill	Chaitrada Chandrama	Chaitrada Chandrama	Amelia Lópes O'Neill
4	Who was born first, Lamine Fomba or Michael Hagan?	Michael Hagan	Lamine Fomba	Lamine Fomba	Michael Hagan
5	Which film has the producer who was born earlier, Term Life or Housefull 2?	Housefull 2	Term Life	Term Life	Housefull 2

Table 3: Questions answered correctly only by SeaKR.

No.	Question	SeaKR ✗	WPSF ✓	CPSF ✗	Gold Answer
6	Which film has the director who is older, The Girl King or Happy As The Grass Was Green?	The Girl King	Happy As The Grass Was Green	The Girl King	Happy As The Grass Was Green
7	Which film has the director who was born first, Nagalit Ang Buwan Sa Haba Ng Gabi or Roberto Carlos E O Diamante Cor-De-Rosa?	Nagalit Ang Buwan Sa Haba Ng Gabi	Roberto Carlos E O Diamante Cor-De-Rosa	Nagalit Ang Buwan Sa Haba Ng Gabi	Roberto Carlos E O Diamante Cor-De-Rosa
8	Which film has the director born later, The Cowboy's Flute or The White Slave (Film)?	The White Slave (Film)	The Cowboy's Flute	The White Slave (Film)	The Cowboy's Flute
9	Where was the place of burial of Ælfgar, Earl Of Mercia's mother?	Oxford	Coventry	Ramsey	Coventry
10	Where was the place of death of the composer of song Love Is Here To Stay?	New York City	Hollywood	New York City	Hollywood

Table 4: Questions answered correctly only by WPSF.

No.	Question	SeaKR ✗	WPSF ✗	CPSF ✓	Gold Answer
11	Which film has the director died earlier, Black Belly Of The Tarantula or Devil On Deck?	Black Belly Of The Tarantula	Black Belly Of The Tarantula	Devil On Deck	Devil On Deck
12	Which film has the director died earlier, Island Of Lost Men or Call Of The Jungle?	Island Of Lost Men	Neither of them	Call Of The Jungle	Call Of The Jungle
13	Which film has the director who died later, The Mystic Masseur or Butter Fingers?	Butter Fingers	17 December 2020	The Mystic Masseur	The Mystic Masseur
14	Which country the director of film Romanoff And Juliet (1961 Film) is from?	England	England	United Kingdom	United Kingdom
15	Where was the place of burial of Karomama I's husband?	Unknown	Luxor	Tanis	Tanis

Table 5: Questions answered correctly only by CPSF.

No.	Question	SeaKR ✗	WPSF ✓	CPSF ✓	Gold Answer
16	Which film has the director who died later, His Last Haul or The Falcon In San Francisco?	His Last Haul	The Falcon In San Francisco	The Falcon In San Francisco	The Falcon In San Francisco
17	Which film has the director died later, Forbidden Dreams or If I Were Free?	If I Were Free	Forbidden Dreams	Forbidden Dreams	Forbidden Dreams
18	Which film has the director who died earlier, The Cruel Truth or The Deadly Companions?	The Deadly Companions	The Cruel Truth	The Cruel Truth	The Cruel Truth
19	Which film has the director died later, Rhythm In The Clouds or September Affair?	September Affair	Rhythm In The Clouds	Rhythm In The Clouds	Rhythm In The Clouds
20	Who died later, Ben Enwonwu or Hezekiah Niles?	Hezekiah Niles	Ben Enwonwu	Ben Enwonwu	Ben Enwonwu

Table 6: Questions answered correctly by both WPSF and CPSF but not by SeaKR.

No.	Question	SeaKR ✗	WPSF ✗	CPSF ✗	Gold Answer
21	What is the date of death of Augusta Leigh's father?	1783	21 May 1791	22nd of August 1811	2 August 1791
22	Where did the director of film Vestire Gli Ignudi die?	Rome	Rome	Rome	Paris
23	Where was the husband of Caterina Visconti born?	Milan	Milan	Milan	Pavia
24	What is the place of birth of the director of film Centaur (2016 Film)?	United States of America	Narnia	Unknown	Kyrgyzstan
25	Which film has the director who was born first, Bhavesh Joshi Superhero or Hellboy II: The Golden Army?	Bhavesh Joshi Superhero	Bhavesh Joshi Superhero	Bhavesh Joshi Superhero	Hellboy II: The Golden Army

Table 7: Questions answered incorrectly by all three runs.