
SPOTLIGHTRAG: ENHANCING FACTUAL ACCURACY WITH POSITION-AWARE SPAN SELECTION

Anonymous authors

Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) enhances LLMs with external knowledge, but current methods face key limitations. Most solutions operate at a coarse passage or sentence level, indiscriminately concatenating retrieved text, which introduces noise, overlooks decisive sub-sentential phrases, and is susceptible to positional bias where evidence is lost in the middle of long contexts. To overcome these challenges, we propose SpotlightRAG, an inference-time framework that enhances factual accuracy through precise, span-level context selection and explicit relevance signaling. SpotlightRAG employs a position-aware scoring mechanism to identify and weight critical text spans, directly countering positional bias. It then uses novel retrieval-aware prefix tokens to explicitly annotate the relevance of each span for the generator, providing fine-grained, interpretable control without model retraining. Extensive experiments on four benchmarks—PopQA, TriviaQA, Natural Questions, and Multi-HopQA—demonstrate that SpotlightRAG consistently outperforms state-of-the-art baselines, including InstructRAG, RankRAG, and In-Context RALM, improving accuracy over strong baselines by 2.1% on PopQA and 1.2% on the challenging MultiHopQA dataset. An anonymized implementation is available at <https://anonymous.4open.science/r/SpotlightRAG-5F6A/>.

1 INTRODUCTION

Large Language Models (LLMs) have significantly transformed information access, enabling users to query vast corpora and obtain fluent, knowledge-grounded responses. However, despite their remarkable capabilities, LLMs remain limited in terms of factual coverage, domain specificity, and knowledge freshness. As a result, hallucinations and factual errors remain common (Huang et al., 2025), particularly in knowledge-intensive tasks such as open-domain question answering (QA). Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm to address these shortcomings by conditioning LLMs on external documents retrieved from large-scale corpora. This design reduces reliance on parametric memorization while improving factual grounding, interpretability, and adaptability.

Nevertheless, the effectiveness of RAG pipelines hinges on how retrieved evidence is represented and integrated. Existing approaches face several critical challenges. A central issue is the trade-off between efficiency and granularity (Khattab & Zaharia, 2020): single-vector dense retrievers often fail to capture fine-grained token alignments and struggle on long-tail queries, whereas multi-vector or late-interaction methods preserve token-level signals but incur substantial storage and computational overhead (Santhanam et al., 2021). Simply extending context length at inference does not resolve this tension; larger retrieval sets frequently introduce irrelevant spans, diluting useful evidence and exacerbating the lost-in-the-middle effect (Yue et al., 2024).

Another limitation lies in coarse integration strategies. Most RAG systems concatenate the top- k passages without discrimination, inevitably introducing redundancy and distractors. Such retrieval noise can significantly reduce accuracy, particularly in multi-hop reasoning where irrelevant content may cause cascading errors (Yu et al., 2024; Yang et al., 2018). Although re-ranking and retriever-aware prompting mitigate this issue to some extent, they largely operate at the passage or sentence level (Li et al., 2024), overlooking sub-sentential cues—such as entities, phrases, or attributive clauses—that often provide decisive evidence (Nematov et al., 2025). Finally, retrieval and generation remain loosely coupled. While some methods jointly pretrain retrievers and generators or refine indexing with similarity modeling, they still offer limited inference-time control. The

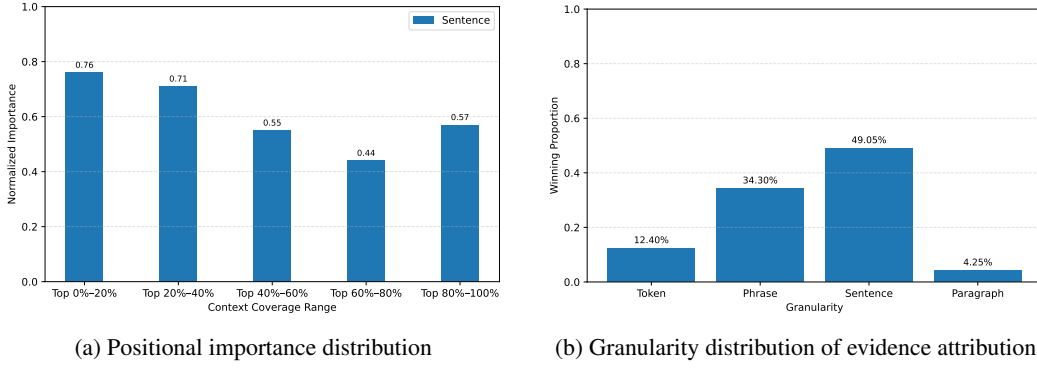


Figure 1: Empirical motivation for fine-grained selection. (a) Positional distribution of evidence importance: strongest in the *first 20%*, gradually declining across the *20–80% middle*, and partially rebounding in the *final 20%*. (b) Granularity distribution: phrase- and sentence-level spans dominate, confirming that sub-sentential scopes are especially diagnostic for RAG.

generator passively consumes retrieved content without explicit mechanisms to amplify informative spans or suppress distractors. This disconnect often leads to diluted attention and brittle reasoning, especially in multi-hop scenarios.

Beyond these challenges, recent studies have highlighted two important trends in the RAG community. On the one hand, there is growing interest in *interpretability and attribution*, with methods seeking to trace answers back to supporting evidence (Goldshmidt & Horovitz, 2024). On the other hand, there is a push toward *fine-grained evidence modeling*, such as span-level attribution, hierarchical retrieval units, and structure-aware selection. Despite these advances, existing frameworks often treat fine-grained signals as auxiliary rather than central components, leaving open the question of how to design a unified pipeline that integrates them directly into inference-time decision making.

To further investigate, we analyze how evidence importance is distributed across document positions. Figure 1 presents two key findings: (a) **positional priors matter**—useful evidence is not evenly distributed but is most concentrated in the *first 20% of sentences*, gradually decreases in the *20–80% middle range*, and shows a partial rebound in the *final 20%*; and (b) **granularity matters**—phrase- and sentence-level spans dominate evidence attribution (Goldshmidt & Horovitz, 2024), while token-only and paragraph-only signals are much less diagnostic.

These empirical findings highlight a fundamental gap: current RAG frameworks lack fine-grained, interpretable, and controllable evidence selection mechanisms. To formalize this, let $\mathcal{R}$ denote the set of retrieved contexts and $\mathcal{R}^*$ the (unknown) set of truly relevant spans. The expected retrieval utility for a query q can be expressed as:

$$U(q) = \mathbb{E}_{s \sim \mathcal{R}} [P(y = 1 \mid q, s)], \quad (1)$$

where $P(y = 1 \mid q, s)$ represents the probability that a span s is relevant to the query. In practice, $U(q)$ degrades rapidly as distractors accumulate, underscoring the need for inference-time selection strategies that approximate $\mathcal{R}^*$ at the span level rather than passively accepting $\mathcal{R}$.

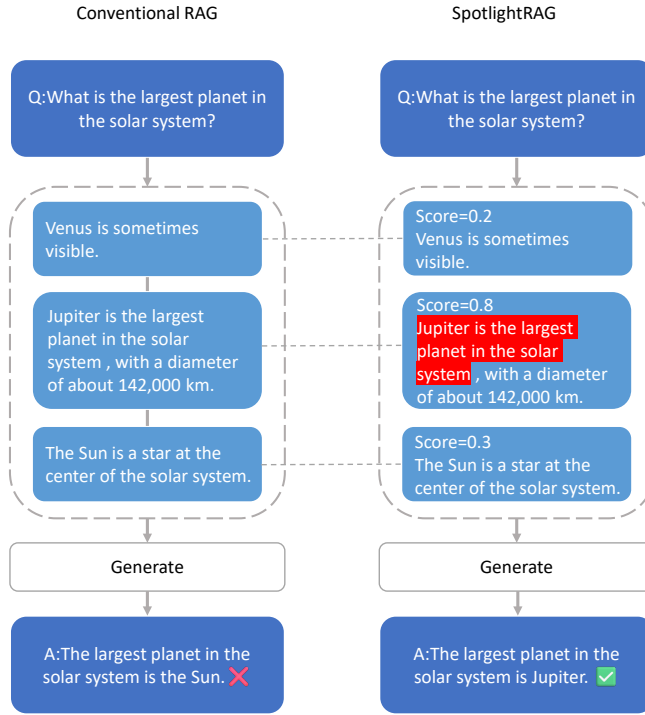


Figure 2: Comparison of conventional RAG and SpotlightRAG. Conventional pipelines often admit distractors (e.g., Sun, Venus), whereas SpotlightRAG emphasizes the decisive span about Jupiter, thereby guiding the generator toward the correct answer.

Motivated by these observations, we propose SpotlightRAG, a fine-grained context selection framework that unifies token-to-phrase alignment, position-aware priors, dynamic relevance weighting, and retrieval-aware prefix tokens. By amplifying informative spans while suppressing distractors, SpotlightRAG achieves robust and interpretable retrieval integration, addressing the persistent retrieval-noise challenge in RAG.

Contributions. This paper advances retrieval-augmented generation by explicitly addressing the gap between coarse retrieval and fine-grained evidence integration. SpotlightRAG is designed as an inference-time framework that unifies empirical analysis, theoretical motivation, and practical design, aiming to enhance factual accuracy while preserving interpretability and efficiency. The key contributions are summarized as follows:

1. **Problem identification:** Critical limitations of existing RAG pipelines are highlighted, including coarse retrieval integration, lack of fine-grained control, and absence of inference-time selection.
2. **Empirical analysis:** Position- and granularity-based studies reveal uneven evidence importance, motivating span-level selection.
3. **Framework design:** SpotlightRAG is introduced, combining (a) span-level modeling, (b) dynamic weighting with positional priors, and (c) retrieval-aware prefix tokens.
4. **Experimental validation:** Extensive experiments on multiple benchmarks demonstrate consistent improvements, confirming both robustness and interpretability.

2 RELATED WORK

2.1 ADVANCES IN RETRIEVAL-AUGMENTED GENERATION

Retrieval-Augmented Generation (RAG) has emerged as a dominant paradigm for enhancing large language models with external knowledge. Early implementations concatenated the top- k retrieved

passages into the model input (Lewis et al., 2021), which improves recall but also introduces redundancy and distractors that dilute model attention. Later research refined this pipeline by developing stronger dense retrievers, hybrid lexical–semantic search, and lightweight re-ranking strategies to suppress obvious noise. However, a persistent challenge remains the trade-off between efficiency and granularity: single-vector dense retrievers struggle to capture fine-grained token alignments, while late-interaction retrievers (Santhanam et al., 2021) preserve token-level signals at high memory cost. Long-context transformers (Tworkowski et al., 2023) extend input windows but often suffer from diminishing returns due to the lost-in-the-middle effect (Liu et al., 2024; Yue et al., 2024).

2.2 FINE-GRAINED EVIDENCE SELECTION AND ATTRIBUTION

Beyond retrieval quality, finer-grained integration has been widely explored. Sentence-level re-ranking improves precision by filtering irrelevant sentences (Yu et al., 2024; Li et al., 2024), but it overlooks decisive sub-sentential cues such as entities, phrases, and appositive clauses. Token-level alignment methods provide higher resolution, though they can be brittle without positional priors and often fail to aggregate signals robustly. Attribution analyses further reveal that phrase- and sentence-level spans dominate evidence importance, while token-only or paragraph-only signals are less diagnostic (Goldshmidt & Horovitz, 2024).

Recent work has more directly addressed these issues. Han et al. (2025) propose Fine-grained Knowledge Enhancement (FKE), which retrieves sentence-level knowledge and constrains decoding with fine-grained signals (Han et al., 2025). Xu et al. (2024) present Tok-RAG, formalizing a theory of benefit vs. detriment at the token level (Xu et al., 2025). Qi et al. (2024) introduce MIRAGE, which attributes generated answers to retrieved documents by analyzing saliency in internal activations (Qi et al., 2024). While these approaches provide finer attribution and improved transparency, they largely operate as post-hoc explanations or decoding constraints rather than offering explicit inference-time control.

2.3 LIMITATIONS AND OPEN CHALLENGES

Despite these advances, existing RAG methods still face critical limitations. Retrieval integration is typically performed at the passage or sentence level, overlooking short spans that carry decisive evidence. Structural heuristics, such as prioritizing lead sentences, capture some positional biases but fail to generalize across contexts. Interpretability also remains limited: exposing retrieved passages or citations shows what was retrieved, but not how much each piece should matter, leaving the generator with little guidance on amplifying or suppressing spans.

Concurrent frameworks such as InstructRAG (Wei et al., 2025), RankRAG (Yu et al., 2024), StructRAG (Li et al., 2024), and more recent designs including REGENT (Sridhar et al., 2025) and MMed-RAG (Xia et al., 2024) refine retrieval through instruction-guided prompting, structured units, or multi-modal reasoning. Yet these remain constrained by coarse granularity and lack explicit inference-time controllability. In contrast, SpotlightRAG addresses this gap by combining token-to-phrase alignment, positional priors, dynamic weighting, and retrieval-aware prefix tokens, thereby enabling robust, fine-grained, and interpretable span selection directly at inference time without retraining.

3 SPOTLIGHTRAG

3.1 TASK FORMULATION

Retrieval-augmented generation can be formalized as follows. Given a query $q = \{q_1, q_2, \dots, q_n\}$, and a retrieved context set $\mathcal{C} = \{c^{(1)}, c^{(2)}, \dots, c^{(M)}\}$, the goal is to generate an answer y conditioned on a compact subset $\mathcal{S}^* \subset \mathcal{C}$ that contains the most relevant spans. Unlike conventional RAG, which concatenates the top- k passages, the proposed formulation explicitly models the expected utility of selected spans:

$$\mathcal{S}^* = \arg \max_{\mathcal{S} \subset \mathcal{C}} \mathbb{E}_{s \in \mathcal{S}} [P(y \mid q, s)], \quad (2)$$

where $P(y \mid q, s)$ denotes the contribution of span s to generating the correct answer. This definition highlights that the quality of selection, rather than the number of retrieved passages, is critical.

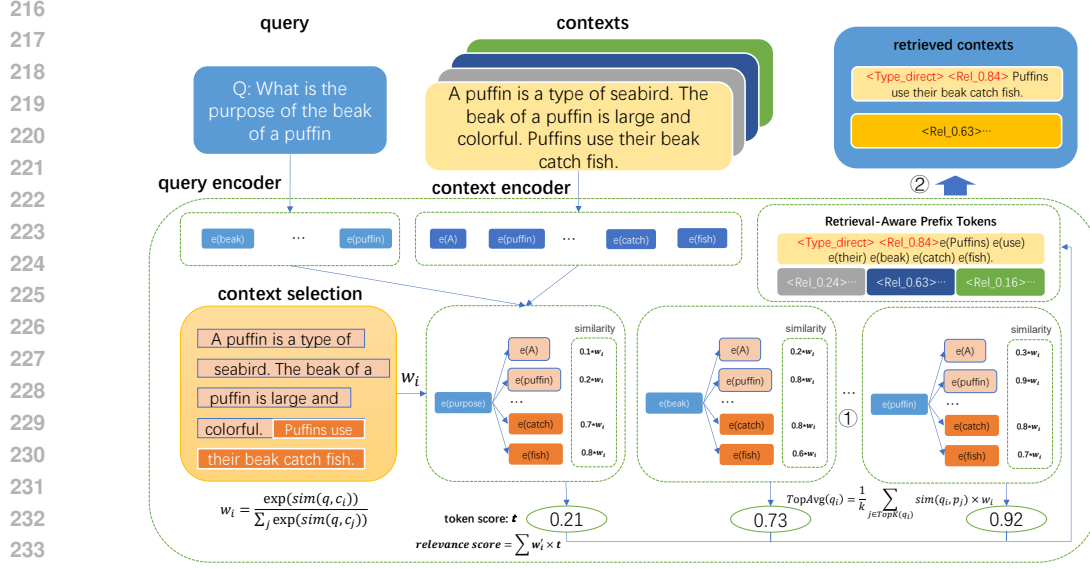


Figure 3: **Overview of the SpotlightRAG pipeline.** Phase 1: fine-grained scoring via token-to-phrase alignment. Phase 2: annotation with score tokens and category tokens. This two-phase integration suppresses distractors and highlights decisive spans.

3.2 FINE-GRAINED EVIDENCE INTEGRATION

The SpotlightRAG framework, illustrated in Figure 3, integrates evidence through two complementary phases that jointly suppress distractors and highlight decisive spans.

Phase 1: Fine-grained scoring. Queries and contexts are first encoded with position-aware embeddings, after which token-to-token similarities are computed. These similarities are aggregated with dynamic relevance weighting to assign scores to spans, ensuring that informative phrases and entities are emphasized even within long passages. This fine-grained scoring mechanism provides a more precise alternative to coarse passage concatenation.

Phase 2: Evidence annotation. The top-ranked sentences derived from Phase 1 are explicitly annotated with two types of prefix tokens: a *score token*, which reflects their computed relevance, and a *category token*, which indicates the type of evidence (Li & Liang, 2021). By injecting these annotations into the input, the generator receives interpretable and controllable signals during inference, enabling span-level integration that goes beyond implicit attention mechanisms.

3.3 TOKEN-LEVEL ENCODING AND RELEVANCE SCORING

Each token from query and context is embedded using a shared encoder with positional priors:

$$\mathbf{h}_{q_i} = \mathcal{E}(q_i) + \mathbf{p}_{q_i}, \quad (3)$$

$$\mathbf{h}_{c_j} = \mathcal{E}(c_j) + \mathbf{p}_{c_j}. \quad (4)$$

Here, $\mathcal{E}(\cdot)$ provides contextualized embeddings while $\mathbf{p}$ encodes sentence or document position, so that early or concluding segments receive stronger inductive bias.

The similarity between each query token q_i and context token c_j is then computed as

$$s_{i,j} = \text{sim}(\mathbf{h}_{q_i}, \mathbf{h}_{c_j}), \quad (5)$$

where cosine similarity is used in practice. To capture global importance, each context token receives a normalized relevance weight:

$$w_j = \frac{\exp(\max_i s_{i,j})}{\sum_k \exp(\max_i s_{i,k})}, \quad (6)$$

which emphasizes tokens highly aligned with at least one query token. This step ensures that even within long passages, fine-grained clues such as entities or attributive phrases can dominate the scoring.

3.4 SENTENCE-LEVEL AGGREGATION

For each query token, only the top- k aligned context tokens are retained, avoiding noise from weak matches. The aggregated relevance score for query token q_i is defined as

$$\text{TopAvg}(q_i) = \frac{1}{k} \sum_{c_j \in \mathcal{T}_k(q_i)} s_{i,j} \cdot w_j, \quad (7)$$

where $\mathcal{T}_k(q_i)$ denotes the top- k aligned tokens. This aggregation balances local similarity with global weighting, stabilizing the contribution of each query token.

At the sentence level, a candidate sentence s_t consisting of tokens $\{c_{j_1}, \dots, c_{j_\ell}\}$ is scored by

$$R(s_t) = \sum_{i=1}^n \alpha_i \cdot \text{TopAvg}(q_i), \quad (8)$$

where α_i represent the importance weights of query tokens. In practice, these weights are not manually assigned but are produced by a lightweight neural scoring module inspired by ColBERT (Khattab & Zaharia, 2020). Specifically, our improved variant computes token-level contributions through late-interaction matching, providing finer attribution of which query terms drive relevance. Importantly, this neural module is independent of the generator and does not require retraining the language model itself. To ensure robustness, we also compare against a uniform weighting scheme ($\alpha_i = 1/n$), which achieves comparable results, confirming that SpotlightRAG remains effective without relying on this additional component. The top- K sentences according to $R(s_t)$ are then selected as the evidence set.

3.5 RETRIEVAL-AWARE PREFIX TOKENS AND INFERENCE

Each selected sentence is annotated with a retrieval-aware prefix token. The literal tag associated with sentence s_t is denoted by its normalized relevance score $R(s_t) \in [0, 1]$ using math-safe notation:

$$s'_t = \langle \text{Rel}_{R(s_t)} \rangle \oplus s_t. \quad (9)$$

Here, $\langle \text{Rel}_{R(s_t)} \rangle$ is a symbolic prefix token whose subscript directly encodes the continuous-valued relevance score, and $\oplus$ denotes concatenation.

In practice, the normalized relevance score $R(s_t) \in [0, 1]$ is directly encoded into the prefix token, resulting in a finite but continuous-looking vocabulary such as $\langle \text{Rel}0.82 \rangle$ or $\langle \text{Rel}0.35 \rangle$. This design avoids the need for coarse binning (e.g., high/medium/low) while still conveying interpretable and fine-grained signals of span importance to the generator.

To further enrich controllability, an additional category token can be attached to indicate the semantic type of evidence (e.g., entity-level, temporal, causal).

By concatenating the query with these annotated sentences, the final input to the generator becomes

$$y = \text{LM}(q \oplus \{s'_t\}_{s_t \in S^*}). \quad (10)$$

Because prefix tokens are appended in text form, this pipeline requires no retraining of the generator and provides transparent inference-time control of how evidence influences generation. Compared with R^2AG (adaptive gating) and RankRAG (sentence-level re-ranking), SpotlightRAG achieves finer granularity, interpretability, and efficiency.

3.6 COMPLEXITY ANALYSIS

SpotlightRAG is designed to remain lightweight at inference time. The overall complexity can be analyzed in terms of time and space:

Time Complexity. For each retrieved context, token-level similarity scoring requires $O(|q| \cdot |c|)$ operations. When multiple contexts are retrieved (e.g., N passages), the overall complexity becomes

$$O(N \cdot |q| \cdot |c|). \quad (11)$$

In practice, N is typically 5–10, and each context is truncated to 512 tokens, yielding manageable cost. Empirically, the overhead relative to a baseline RAG pipeline remains below 5% even when processing $N = 10$ contexts end-to-end on a single A100 GPU. Detailed latency results for $N = 1, 5, 10$ are provided in the appendix, confirming that scaling to realistic retrieval settings does not alter the overall efficiency claim.

Space Complexity. The memory footprint is dominated by the similarity matrix between query and context tokens, which has size $O(|q| \cdot |c|)$. Since only the top- k alignments are retained for each query token, the effective storage is reduced to $O(|q| \cdot k)$. Additional space for storing annotated prefix tokens is linear in the number of selected sentences, which is typically small. Hence the overall space complexity remains manageable.

In summary, SpotlightRAG achieves fine-grained evidence selection with computational and memory costs that are comparable to standard RAG pipelines, ensuring theoretical scalability without introducing additional model parameters.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION METRICS

The proposed framework is evaluated on four widely used open-domain QA benchmarks: **PopQA** (Mallen et al., 2023), **TriviaQA** (Joshi et al., 2017), **Natural Questions** (Kwiatkowski et al., 2019), and **MultiHopQA** (Schnitzler et al., 2024). Each dataset is split into training, validation, and test subsets following standard practice. For evaluation, the Exact Match (EM) Accuracy metric is adopted, which measures whether the generated answer exactly matches the ground truth. This metric has been widely used in prior RAG studies (e.g., InstructRAG, RankRAG, RetRobust) and provides a stringent indicator of factual correctness.

4.2 BASELINES

The proposed method is compared against several competitive RAG-based baselines:

- **InstructRAG (ICLR 2025)** (Wei et al., 2025): Enhances relevance by combining retrieval with instruction-guided prompting.
- **RetRobust (ICLR 2024)** (Yoran et al., 2024): Improves generalization via robust retrieval and noise-aware re-ranking.
- **Self-RAG (ICLR 2024)** (Asai et al., 2023): Introduces self-reflection during retrieval and generation.
- **R^2AG (arXiv 2024)** (Ye et al., 2024): Incorporates re-ranking and adaptive gating for retrieval refinement.
- **RankRAG (NeurIPS 2024)** (Yu et al., 2024): Employs sentence-level re-ranking to filter retrieval noise.
- **In-Context RALM (TACL 2023)** (Ram et al., 2023): Leverages retrieval-augmented in-context learning by injecting retrieved rationales into prompts, improving controllability and factual grounding.

To ensure a fair comparison, the implementation uses the same backbone retriever and generator as the baselines.

4.3 EXACT MATCH ACCURACY RESULTS

For all main experiments, we set the number of selected sentences $K = 6$ and the token-level Top- $k = 5$, as these values demonstrated optimal performance in our sensitivity analysis (see Table 4).

Table 1 and Table 2 summarize the performance across datasets and training configurations. Results are reported under two retrieval settings: w/ Training, where retrieval models are jointly trained with the generator, and w/o Training, where frozen retrievers are used.

As shown in Table 1, the proposed method consistently achieves the best Exact Match scores on both PopQA and TriviaQA. On PopQA, SpotlightRAG surpasses InstructRAG by +2.1% with training and +1.4% without training. On TriviaQA, the framework slightly improves over RankRAG and InstructRAG, reaching 79.1% EM with training and 82.7% without training. These results indicate that the fine-grained scoring mechanism contributes to stable gains across single-hop QA benchmarks.

In addition, Table 2 reports results on Natural Questions and the more challenging MultiHopQA benchmark. On Natural Questions, SpotlightRAG achieves the highest accuracy (66.9% with training, 64.2% without training), outperforming InstructRAG, RetRobust, and RankRAG. On MultiHopQA, the method also leads by a clear margin, reaching 58.4% with training and 51.3% without

Table 1: Exact Match Accuracy (%) on PopQA, TriviaQA benchmarks.

Method	Source	PopQA		TriviaQA	
		w/Train	w/o Train	w/Train	w/o Train
InstructRAG	ICLR 2025	66.2	65.5	78.5	81.2
RetRobust	ICLR 2024	56.5	53.9	71.5	67.2
Self-RAG	ICLR 2024	55.8	52.7	71.4	69.3
R^2AG	arXiv 2024	64.4	65.3	74.2	73.5
RankRAG	NeurIPS 2024	64.1	63.8	78.8	81.9
In-Context RALM	TACL 2023	62.3	61.2	70.2	71.4
Ours	This paper	68.3	66.9	79.1	82.7

Table 2: Exact Match Accuracy (%) on MultiHopQA and NaturalQuestions benchmark.

Method	Source	NaturalQuestions		MultiHopQA	
		w/Train	w/o Train	w/Train	w/o Train
InstructRAG	ICLR 2025	65.7	62.1	57.2	50.4
RetRobust	ICLR 2024	54.2	51.3	53.4	49.2
Self-RAG	ICLR 2024	42.8	40.9	32.9	30.1
R^2AG	arXiv 2024	66.3	63.2	53.2	49.8
RankRAG	NeurIPS 2024	53.2	50.6	37.2	35.3
In-Context RALM	TACL 2023	52.3	50.7	45.2	43.4
Ours	This paper	66.9	64.2	58.4	51.3

training, exceeding InstructRAG, RankRAG, and In-Context RALM. These improvements across both single-hop and multi-hop QA tasks highlight the robustness and generality of the proposed approach.

4.4 ABLATION STUDY

To better understand the contribution of each component in the proposed framework, an ablation study is conducted. Starting from the full model, individual modules are removed, including the fine-grained scorer, importance weighting, and special tokens.

As shown in Table 3, each module contributes positively to the final performance. Removing the **fine-grained scorer** causes the most substantial drop, reducing EM by up to 3% on PopQA and 3.7% on TriviaQA, confirming its critical role in capturing token-level alignment. Excluding the **importance weighting** module also reduces accuracy, particularly on TriviaQA. The absence of **special tokens** leads to a modest but consistent decline, showing that retrieval-aware signals benefit the generator.

These results highlight that dynamic scoring and selective integration jointly enhance factual recall and robustness under retrieval noise. Together with the ablation results, these observations indicate that each component of SpotlightRAG plays a complementary role, and the framework as a whole achieves strong factual accuracy, robustness to noise, and resilience under parameter variation.

Table 3: Ablation study on PopQA, TriviaQA, and Natural Questions under trained retrieval. Removing any component decreases performance, while the full SpotlightRAG achieves the best accuracy.

Model Variant	PopQA	TriviaQA	NaturalQuestions	MultiHopQA
SpotlightRAG	68.3	79.1	66.9	58.4
w/o Fine-grained Scorer	65.9	75.4	64.5	50.2
w/o Importance Weighting	66.1	76.2	64.9	55.6
w/o Special Tokens	66.4	77.2	65.3	52.7

Table 4: Sensitivity analysis of SpotlightRAG under different parameter settings. Accuracy (Acc.) values are left blank (–) for later completion.

Hyper-parameter Setting	Accuracy (%)
Sentence Selection $K = 2$	66.9
Sentence Selection $K = 4$	67.8
Sentence Selection $K = 6$	68.3
Sentence Selection $K = 8$	67.1
Token-level Top- $k = 1$	67.3
Token-level Top- $k = 3$	67.6
Token-level Top- $k = 5$	68.3
Token-level Top- $k = 7$	67.2
Positional Prior Weight = 0.0 (No bias)	65.1
Positional Prior Weight = 0.5 (Moderate)	68.3
Positional Prior Weight = 1.0 (Strong)	67.9
Context Length = 512 tokens	67.7
Context Length = 1024 tokens	68.0
Context Length = 2048 tokens	68.3

4.5 HYPER-PARAMETER SENSITIVITY STUDY

In addition to the ablation study, sensitivity to key hyperparameters is further investigated, including the number of selected sentences K , the token-level Top- k parameter used in relevance aggregation, and the weighting coefficient for positional priors. These parameters directly influence the balance between precision and recall in evidence selection.

First, the number of selected sentences K is varied to examine how much context the generator requires for optimal performance. Results show that performance improves when increasing K from very small values (e.g., 1–2 sentences) to moderate values (e.g., 4–6 sentences), but plateaus or slightly degrades when K is too large, as irrelevant sentences begin to dilute attention. This confirms that SpotlightRAG benefits from compact but informative evidence sets.

Next, different values of the token-level Top- k parameter in the aggregation function are tested. Smaller values emphasize only the strongest alignments, while larger values incorporate broader contextual matches. The results indicate that moderate values (e.g., $k = 3$ or 5) achieve the best trade-off, capturing sufficient evidence without introducing excessive noise.

Finally, the influence of positional prior weighting is assessed. Without positional priors, the model tends to overemphasize mid-passage tokens, leading to reduced accuracy. With appropriately scaled priors, SpotlightRAG consistently identifies lead and concluding spans as more informative, which aligns with the empirical findings in Figure 1.

Overall, the sensitivity analysis validates that SpotlightRAG remains stable across a wide range of hyperparameter values. While performance peaks at moderate settings, even suboptimal choices do not cause severe degradation, demonstrating the practicality of the design in real-world deployments where extensive tuning is infeasible.

5 CONCLUSION

This work presents SpotlightRAG, a fine-grained evidence selection framework for retrieval-augmented generation (RAG). The framework introduces three key innovations: span-level modeling that pinpoints decisive phrases beyond sentence-level filtering, position-aware dynamic weighting that alleviates the lost-in-the-middle effect, and retrieval-aware prefix tokens that provide explicit and controllable guidance during inference. Together, these components enable lightweight yet effective inference-time integration without retraining. Experiments across single-hop and multi-hop QA benchmarks demonstrate consistent gains over strong baselines, highlighting both robustness and interpretability. Looking ahead, the framework offers a promising foundation for extending fine-grained evidence selection to broader settings such as cross-domain adaptation, long-context reasoning, and multimodal RAG applications (Gao et al., 2024).

ETHICS STATEMENT

This research does not involve human subjects, personal data, or sensitive demographic attributes. All experiments are conducted on publicly available datasets (PopQA, TriviaQA, Natural Questions, and MultiHopQA), which are established benchmarks in open-domain question answering research. Potential ethical concerns in retrieval-augmented generation include the risk of propagating biased or factually incorrect content from retrieved documents. The proposed framework aims to mitigate these risks by improving factual accuracy, interpretability, and robustness of retrieval integration. Code and evaluation setups will be made available to facilitate transparency and accountability.

REPRODUCIBILITY STATEMENT

Reproducibility is ensured through the use of publicly available datasets, which are properly cited. The implementation builds on standard open-source frameworks, and the complete source code, including training and evaluation scripts, span selection modules, and configuration files, will be released upon publication. Detailed hyperparameter settings, evaluation protocols, and results of ablation and sensitivity studies are provided in the main text and appendix. These resources enable independent verification and further extension of the reported results by the research community.

REFERENCES

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023. URL <https://arxiv.org/abs/2310.11511>.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL <https://arxiv.org/abs/2312.10997>.
- Roni Goldshmidt and Miriam Horovicz. Tokenshap: Interpreting large language models with monte carlo shapley value estimation, 2024. URL <https://arxiv.org/abs/2407.10114>.
- Jingxuan Han, Zhendong Mao, Yi Liu, Yexuan Che, Zheren Fu, and Quan Wang. Fine-grained knowledge enhancement for retrieval-augmented generation. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL <https://api.semanticscholar.org/CorpusID:280034504>.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL <http://dx.doi.org/10.1145/3703155>.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147/>.
- O. Khattab and Matei A. Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. URL <https://api.semanticscholar.org/CorpusID:216553223>.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl.a.00276. URL <https://aclanthology.org/Q19-1026/>.

-
- 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. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL <https://arxiv.org/abs/2005.11401>.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021. URL <https://arxiv.org/abs/2101.00190>.
- Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. *arXiv preprint arXiv:2410.08815*, 2024.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a.00638. URL <https://aclanthology.org/2024.tacl-1.9/>.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, 2023. URL <https://arxiv.org/abs/2212.10511>.
- Ikhtiyor Nematov, Tarik Kalai, Elizaveta Kuzmenko, Gabriele Fugagnoli, Dimitris Sacharidis, Katja Hose, and Tomer Sagi. Source attribution in retrieval-augmented generation, 2025. URL <https://arxiv.org/abs/2507.04480>.
- Jirui Qi, Gabriele Sarti, Raquel Fern andez, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6037–6053. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.347. URL <http://dx.doi.org/10.18653/v1/2024.emnlp-main.347>.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models, 2023. URL <https://arxiv.org/abs/2302.00083>.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*, 2021.
- Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. Morehopqa: More than multi-hop reasoning, 2024. URL <https://arxiv.org/abs/2406.13397>.
- Kaustubh Sridhar, Souradeep Dutta, Dinesh Jayaraman, and Insup Lee. Regent: A retrieval-augmented generalist agent that can act in-context in new environments, 2025. URL <https://arxiv.org/abs/2412.04759>.
- Szymon Tworkowski, Konrad Staniszewski, Miko aj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Mi o . Focused transformer: Contrastive training for context scaling, 2023. URL <https://arxiv.org/abs/2307.03170>.
- Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=P1qhkp8gQT>.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024.
- Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. A theory for token-level harmonization in retrieval-augmented generation, 2025. URL <https://arxiv.org/abs/2406.00944>.

-
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL <https://arxiv.org/abs/1809.09600>.
- Fuda Ye, Shuangyin Li, Yongqi Zhang, and Lei Chen. R²ag: Incorporating retrieval information into retrieval augmented generation, 2024. URL <https://arxiv.org/abs/2406.13249>.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context, 2024. URL <https://arxiv.org/abs/2310.01558>.
- Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms, 2024. URL <https://arxiv.org/abs/2407.02485>.
- Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. *arXiv preprint arXiv:2410.04343*, 2024.

USE OF LLMs

Large Language Models (LLMs) were used solely for spelling correction and English grammar checking in the preparation of this manuscript. No LLM was involved in generating research ideas, methods, analyses, or experimental results.