applied sciences

MDPI

Article

SES-RAG: An Adaptive Framework with ModernBERT Segmentation and Generative-Alignment Expansion

Chaobang Gao, Hao He and Qiang Zhang *

School of Computer Science, Chengdu University, Chengdu 610106, China; hehao@cdu.edu.cn (H.H.)
* Correspondence: cduzhangqiang@163.com

Abstract

Retrieval-augmented generation (RAG) is promising for long-document question answering, but its performance is often limited by fixed-window chunking, query-document mismatch, and static context truncation. This study proposes SES-RAG, a lightweight framework that jointly improves semantic segmentation, retrieval alignment, and context selection in long-document RAG. Specifically, SES-RAG includes a ModernBERT-based semantic segmenter for constructing semantically coherent chunks, Generative-Alignment Query Expansion (GAQE) for enhancing query-document alignment through residual fusion with a hypothetical document, and Information Density Truncation (IDT) for adaptively selecting retrieved contexts. We evaluate SES-RAG on three long-document QA benchmarks, namely NarrativeQA, QuALITY, and QASPER, against Naive RAG, HyDE-RAG, Hybrid-RAG, and CRAG. SES-RAG achieves the best overall performance on all three datasets, reaching 15.71 ROUGE, 9.61 BLEU-1, 6.16 BLEU-4, and 18.71 METEOR on NarrativeQA, 74.33 accuracy on QuALITY, and 41.06 F1-Match on QASPER. Compared with Naive RAG, SES-RAG improves ROUGE by 5.10 points on NarrativeQA, Accuracy by 3.66 points on QuALITY, and F1-Match by 3.83 points on QASPER. These results show that lightweight cross-stage optimization is an effective and practical strategy for improving long-document RAG.

Keywords: large language models; long-document question answering; ModernBERT; query expansion; retrieval-augmented generation; semantic segmentation; adaptive context selection

1. Introduction

As large language models (LLMs) continue to advance natural language processing, addressing their inherent limitations—especially knowledge cutoff and hallucination—has become a central research focus. Retrieval-augmented generation (RAG) provides an effective non-parametric solution by introducing external knowledge sources at inference time, thereby improving factual grounding and adaptability in knowledge-intensive tasks [1]. However, although RAG has shown strong performance in open-domain question answering and long-document understanding, standard pipelines remain fragile when handling long documents, where relevant evidence is often scattered across distant passages and embedded in complex discourse structures [2].

A major challenge is that the effectiveness of long-document RAG depends not only on retrieval itself, but also on how semantic information is preserved and propagated across the full pipeline. In practice, current RAG systems still suffer from three recurring bottlenecks. First, semantic fragmentation arises when documents are split by fixed windows or

Check for updates

Academic Editor: George Drosatos

Received: 17 February 2026
Revised: 17 April 2026
Accepted: 22 April 2026
Published: 13 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

simple syntactic rules, which often break coherent discourse units and damage evidence completeness. Second, the query–document modality gap remains a persistent obstacle: user queries are usually short interrogative expressions, whereas source documents are long declarative passages, making it difficult for embedding-based retrieval to capture true semantic relevance. Third, signal-to-noise imbalance caused by static truncation limits context utilization efficiency. Fixed Top-K selection may retain too many irrelevant passages in sparse-relevance cases, while truncating useful long-tail evidence in dense-relevance cases.

Recent studies have attempted to strengthen RAG from different perspectives. Some methods focus on improving segmentation or chunk construction so that retrieval units better preserve semantic integrity. Others aim to enhance retrieval through query reformulation, hypothetical-document generation, or hybrid retrieval strategies that combine multiple retrieval signals. Another line of work improves post-retrieval context construction through compression, reranking, or filtering [3]. These approaches are often effective, but most of them optimize only a single stage of the pipeline. As a result, they only partially alleviate the cascading errors that arise when semantic damage introduced during chunking, retrieval mismatch, and noisy context selection interact with each other in long-document settings.

More recently, stronger RAG frameworks have introduced corrective retrieval, iterative refinement, or reflection-based mechanisms to further improve robustness [4]. While these approaches can enhance answer quality, they often require more complex control flows, repeated model calls, or additional retrieval and reranking operations, resulting in higher latency and deployment cost. For practical long-document question answering systems, there remains a clear need for a lightweight framework that can improve multiple critical stages of the pipeline without relying on heavy iterative procedures.

To address this gap, we propose SES-RAG, a semantics-enhanced and adaptive optimization framework for long-document RAG. Rather than introducing a heavy multi-round architecture, SES-RAG performs coordinated optimization over three key stages of the standard pipeline. First, to mitigate semantic damage caused by mechanical chunking, we design a lightweight ModernBERT-based semantic segmenter that detects semantic boundaries and constructs more coherent retrieval units. Second, to reduce query–document expression mismatch, we propose Generative-Alignment Query Expansion (GAQE), which generates a context-rich hypothetical document and aligns it with the original query representation through a residual weighted fusion strategy. Third, we introduce Information Density Truncation (IDT), a training-free adaptive selection mechanism that determines the truncation point of retrieved contexts according to score-distribution-based information density, thereby suppressing long-tail noise while preserving core evidence.

This design is motivated by a simple observation: in long-document RAG, answer quality is determined not only by whether relevant passages are retrieved, but also by whether retrieval units remain semantically intact, whether the query is properly aligned with document semantics, and whether the final context contains an appropriate amount of useful evidence. SES-RAG therefore aims to optimize the semantic flow of the entire retrieval-generation process while remaining modular, practical, and easy to integrate into standard RAG systems.

To validate the proposed framework, we conduct experiments on three long-document QA benchmarks: NarrativeQA, QuALITY, and QASPER. In addition to Naive RAG, we compare SES-RAG against representative baselines including HyDE-RAG, Hybrid-RAG, and CRAG. The results show that SES-RAG consistently achieves the best overall performance across benchmarks. We further perform ablation studies to analyze the contribution of each module, cross-retriever experiments to examine robustness under different retrieval backbones, and efficiency analysis to evaluate the practical overhead of the framework.

The experimental findings indicate that the gains of SES-RAG come not from a single isolated enhancement, but from coordinated improvements in semantic segmentation, query alignment, and adaptive context selection.

The main contributions of this paper are summarized as follows:

1. We propose SES-RAG, a semantics-enhanced and adaptive optimization framework for long-document RAG that jointly improves the index construction, retrieval, and context filtering stages.
2. We introduce three complementary modules: a ModernBERT-based semantic segmenter for semantically coherent chunk construction [5], GAQE for query–document alignment via hypothetical-document fusion, and IDT for adaptive context truncation based on information density.
3. We provide comprehensive empirical evaluation on NarrativeQA, QuALITY, and QASPER, together with ablation and cross-retriever analyses, demonstrating that SES-RAG consistently outperforms strong baseline frameworks while remaining lightweight and practical.

2. Related Work

Related work on long-document RAG can be broadly organized into three directions: text segmentation for index construction, retrieval enhancement through query reformulation or expansion, and post-retrieval context optimization. These directions correspond to different stages of the retrieval pipeline and are closely related to the design of SES-RAG. Existing studies have substantially improved long-document RAG, but many of them focus on only one stage of the pipeline. More recent frameworks further improve robustness through corrective retrieval, iterative refinement, or reflective control, although such designs often introduce additional system complexity and runtime overhead.

2.1. Text Segmentation

Text segmentation is a fundamental step in retrieval index construction. Its purpose is to divide long documents into semantically coherent units so that retrieval operates over complete evidence fragments rather than mechanically truncated text [6]. As shown in Figure 1, traditional chunking strategies usually rely on fixed windows, paragraph boundaries, punctuation rules, or recursive splitting. Although these methods are computationally efficient, they often ignore discourse continuity and may separate logically connected content, such as causal relations, coreference chains, or elaborative explanations, thereby weakening downstream retrieval quality.

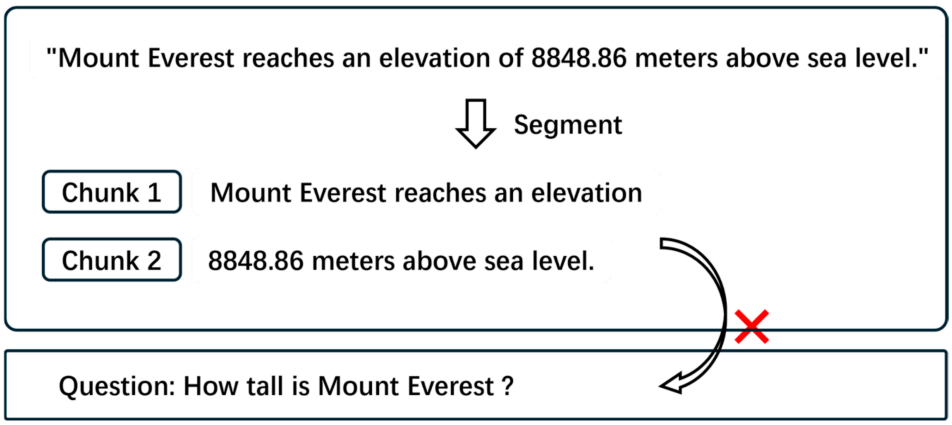


Figure 1. Illustration of semantic fragmentation caused by fixed-window chunking.

Earlier research on text segmentation explored both rule-based and learning-based approaches. Classical methods such as topic-boundary detection identify segmentation points through local lexical cohesion or similarity shifts. Later studies formulate segmentation as a supervised boundary classification problem, in which sequence models or pre-trained encoders determine whether adjacent sentences belong to the same semantic unit. Compared with purely heuristic chunking, these learning-based methods generally provide better boundary consistency and are more compatible with discourse structure, making them more suitable for tasks that require fine-grained evidence preservation.

More recently, LLM-based segmentation has also been investigated [7]. Through zero-shot or few-shot prompting, LLMs can identify semantic breakpoints or iteratively segment long documents based on discourse-level understanding. These methods are attractive in terms of semantic sensitivity, but they often incur substantial latency and API cost, especially when long contexts must be processed repeatedly. For practical long-document RAG systems, a lighter alternative is often more appropriate [8]. In this context, SES-RAG adopts a ModernBERT-based semantic segmenter as a discriminative boundary detector, aiming to improve the semantic completeness of retrieval units while remaining suitable for offline batch processing and large-scale indexing.

2.2. Retrieval and Query Expansion

The mismatch between short user queries and long declarative documents is a central challenge in retrieval-augmented generation. Traditional sparse retrieval methods, such as BM25, rely primarily on lexical matching and usually perform well when query terms explicitly overlap with document text. However, their effectiveness often declines under paraphrasing, abstraction, or domain-specific variation in expression. Dense retrieval alleviates this limitation by mapping queries and documents into a shared semantic vector space, thereby providing stronger generalization across lexical variation. Nevertheless, dense retrieval can still be unstable when the query is underspecified or when semantic alignment between the question and the supporting evidence remains weak [9].

To address this problem, a large body of work focuses on query reformulation and expansion. Early approaches extend the original query through keyword expansion, pseudo-relevance feedback, or concept augmentation in order to improve recall. More recent LLM-based methods generate richer query representations in document-like form. A representative example is HyDE, which generates a hypothetical document and uses it as a retrieval anchor, often improving semantic recall when lexical overlap is limited [10]. In addition, hybrid retrieval strategies combine sparse and dense signals to improve both precision and recall, and have become strong practical baselines in many RAG systems.

At the framework level, some recent methods extend beyond one-shot retrieval by introducing correction, reflection, or retrieval control [11,12]. Corrective retrieval frameworks, for example, attempt to identify low-quality retrieval results and trigger additional retrieval or evidence adjustment steps [13]. These designs can improve robustness, but they also increase runtime cost and overall system complexity. Against this background, GAQE is designed as a lightweight alternative that preserves the benefits of generative expansion while reducing over-reliance on generated content. Inspired by the document-form alignment idea used in HyDE, GAQE generates a context-rich hypothetical document and fuses it with the original query representation through residual weighted alignment, thereby improving semantic matching while preserving the retrieval direction of the original question.

2.3. Context Optimization

With the increasing context window of modern LLMs, post-retrieval context construction has become a crucial factor in end-to-end answer quality. The most common strategy

is fixed Top-K truncation, in which the top-ranked retrieved chunks are directly passed to the generator. In practice, however, retrieval-score distributions are highly variable. As shown in Figure 2, in some cases, only a few passages contain useful evidence and the remaining candidates are mostly noise; in other cases, relevant evidence is distributed across many moderately ranked chunks. Under such conditions, a fixed Top-K policy can either introduce excessive distractor context or prematurely remove useful long-tail evidence [14].

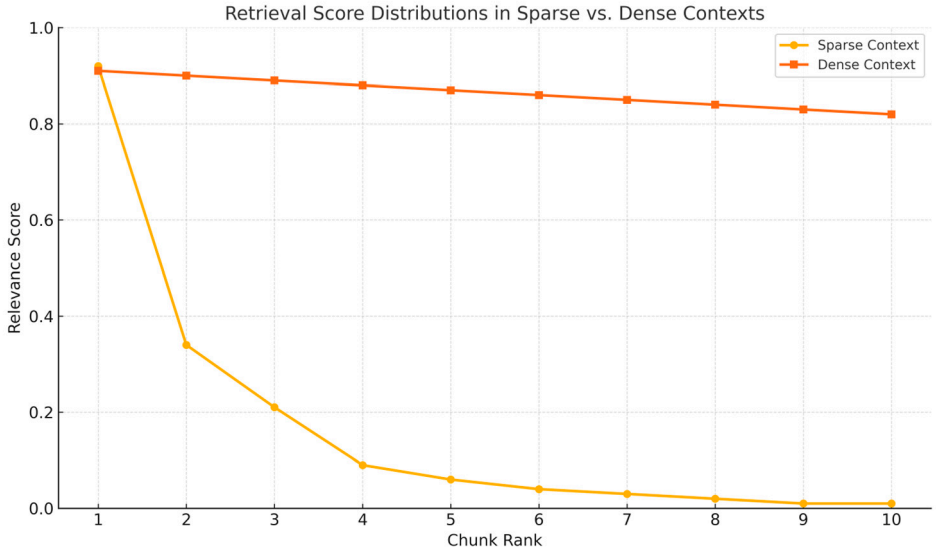


Figure 2. Retrieval Score Distribution in Sparse vs. Dense Contexts.

Several lines of research attempt to address this issue. One direction uses reranking models, such as cross-encoders, to rescore candidate passages more precisely [15]. Although effective, reranking often introduces additional inference overhead and can become expensive when the candidate pool is large. Another direction focuses on prompt compression or context compression, where retrieved contexts are filtered, reordered, or compressed before generation. Recent methods such as LongLLMLingua show that careful context compression can improve generation efficiency and reduce the negative effect of noisy long inputs [16]. At the same time, studies on long-context LLM behavior have shown that simply providing more text does not necessarily lead to better performance; excessively long contexts may dilute key evidence and worsen answer generation.

More recent RAG frameworks may further combine context selection with multi-stage retrieval, corrective refinement, or reflection-based mechanisms. While these approaches can improve robustness, they also tend to rely on additional model calls and more complex orchestration. In contrast, IDT is designed as a low-cost and training-free adaptive context selection strategy. Instead of using a fixed Top-K cutoff, it estimates the truncation point according to the information density implied by the retrieval-score distribution, allowing the system to retain high-value evidence while suppressing low-utility long-tail noise. This design is better aligned with the practical requirements of long-document RAG, where context quality often matters more than simply increasing context quantity.

2.4. Research Gap

Existing studies have shown that text segmentation, retrieval enhancement, and context optimization are all critical to long-document RAG. However, these components are often improved in isolation. As a result, long-document RAG systems may still suffer from cascading weaknesses across the pipeline: mechanically fragmented chunks reduce evidence completeness, query-document mismatch weakens retrieval alignment, and static

truncation may either preserve distractor noise or discard useful long-tail evidence. Although recent corrective or reflective RAG frameworks can improve robustness, they often do so at the cost of additional model calls, higher latency, and greater system complexity. These observations indicate the need for a lightweight and modular framework that can jointly improve semantic chunk construction, retrieval alignment, and adaptive context selection within a standard RAG pipeline. SES-RAG is proposed to address this need.

3. SES-RAG Framework

This section details the design principles and implementation of SES-RAG. The framework consists of three cascaded modules responsible for semantic segmentation, query expansion, and adaptive selection. We first formalize the RAG task, then introduce ModernBERT-based semantic segmentation, generative-alignment query expansion, and information-density based adaptive truncation.

3.1. Framework Overview

The overall pipeline is shown in Figure 3.

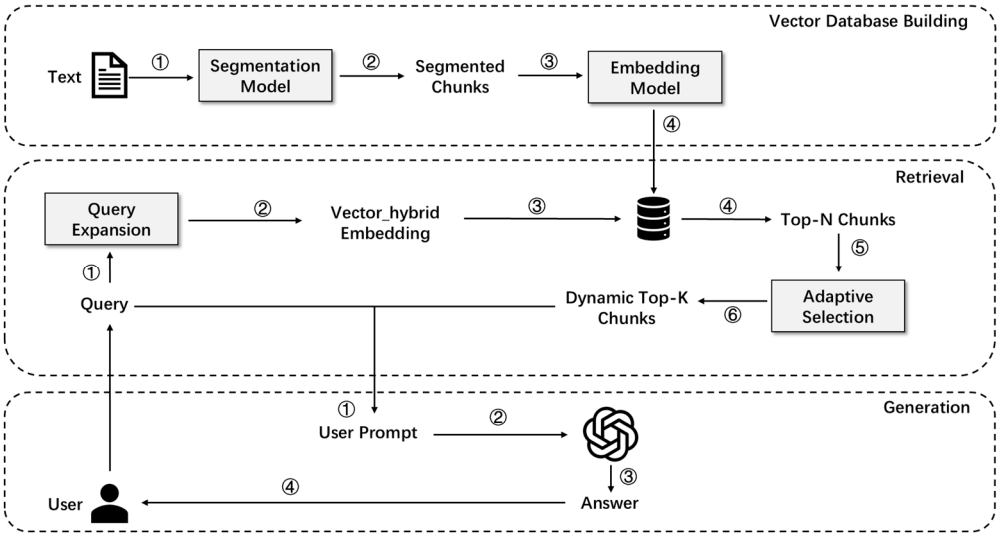


Figure 3. The overall architecture of SES-RAG framework.

Given a user query q and an external document corpus $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$, RAG aims to retrieve the most relevant context set $\mathcal{C}$ and assist an LLM in generating an answer a. SES-RAG optimizes this process through three stages to address semantic fragmentation, modality gap, and signal-to-noise imbalance. During index construction, a discriminative segmenter M_{seg} partitions each long document d into a set of semantically coherent chunks $C = \{c_1, \ldots, c_n\}$, ensuring each chunk is an independent semantic unit. During retrieval, a generator G produces a hypothetical document $\hat{d}$ and performs weighted fusion with the original query q in vector space to obtain an enhanced query embedding v_{hybrid}. Finally, based on the retrieval-score sequence $S = \{s_1, \ldots, s_N\}$, an information-density function $D(k)$ is used to dynamically determine the optimal truncation threshold k^*, producing the final selected context $\mathcal{C}^*$.

3.2. Semantic Segmentation with ModernBERT

Fixed-window chunking in long documents often breaks discourse continuity, coreference chains, and semantically related evidence, resulting in fragmented retrieval units. As illustrated in Figure 4, mechanically split chunks may damage semantic completeness, whereas semantic re-segmentation aims to reorganize the text into more coherent evidence

units. To achieve this in a lightweight and scalable manner, we formulate segmentation as an adjacent-sentence boundary classification task rather than an autoregressive generation problem. This discriminative formulation is computationally lightweight and more suitable for large-scale document indexing than LLM-based generative segmentation.

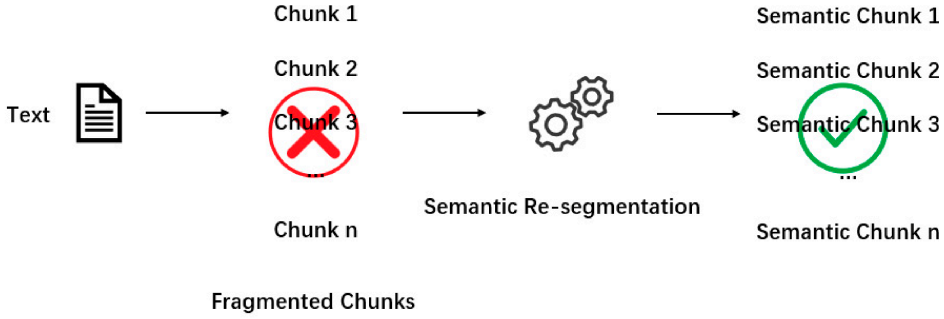


Figure 4. From fragmented chunks to semantically coherent chunks.

3.2.1. Problem Definition

Given a document d, we first apply a rule-based preprocessing step driven by sentence-ending punctuation to obtain an atomic sentence sequence $S = \{sent_1, sent_2, \ldots, sent_L\}$. For any adjacent sentence pair $(sent_i, sent_{i+1})$, the model predicts a binary label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates a semantic shift at this position (i.e., a boundary should be inserted), and $y_i = 0$ indicates semantic continuity (i.e., the two sentences should be merged). Consequently, the segmentation of the entire document is uniquely determined by the boundary sequence $\{y_i\}_{i=1}^{L-1}$: when $y_i = 1$, a split is made between $sent_i$ and $sent_{i+1}$; otherwise, the two sentences are assigned to the same segment and the merging process continues forward [17].

Based on this formulation, semantic segmentation reduces to a lightweight boundary decision problem over adjacent sentence pairs. The input for each decision point is therefore a paired sentence representation, and the output is the boundary probability $P(y_i = 1 \mid x_i)$. We instantiate this classifier with ModernBERT in order to combine long-context modeling ability with efficient discriminative inference.

3.2.2. Model Architecture

We use ModernBERT-base as the backbone encoder for boundary classification. Compared with Bidirectional Encoder Representations from Transformers (BERT), ModernBERT introduces Rotary Positional Embeddings (RoPE) [18], natively supports sequences up to 8192 tokens, and is therefore better suited to modeling long-range discourse continuity and coreference patterns. In addition, its integrated Flash Attention improves long-sequence efficiency and reduces memory overhead [19], making it practical for fine-grained boundary prediction in long-document settings.

For the i-th sentence pair, we construct a paired-text input:

$$x_i = [\text{CLS}] \oplus sent_i \oplus [\text{SEP}] \oplus sent_{i+1} \oplus [\text{SEP}] \tag{1}$$

where $\oplus$ denotes concatenation. The encoder outputs a pair representation:

$$h_i = \text{ModernBERT}(x_i) \tag{2}$$

And a classification head produces the boundary probability:

$$P(y_i = 1 | x_i) = \sigma(W \cdot h_i + b) \tag{3}$$

where σ is the sigmoid function and W, b are head parameters.

3.2.3. Training and Inference

We treat each adjacent pair $(sent_i, sent_{i+1})$ as a training example. The training corpus is constructed from WikiText by first splitting each document into sentences and then forming adjacent sentence pairs. Supervision is derived from human/heuristic-aligned paragraph boundaries: if the two sentences cross a semantic paragraph boundary, we assign label 1; otherwise, label 0. This formulation turns semantic segmentation into a lightweight boundary classification task over adjacent sentence pairs. To mitigate class imbalance, random down-sampling or loss reweighting can be used. In our implementation, the semantic segmenter is trained for 3 epochs using BF16 mixed precision, with the maximum sequence length set to 8192.

At inference time, we compute boundary probabilities for all adjacent pairs. Given a threshold τ, segmentation is performed with a linear merging scan: starting from $sent_1$, if $P(y_i = 1) < \tau$, $sent_{i+1}$ is merged into the current chunk; if $P(y_i = 1) \geq \tau$, a new chunk is started. In our implementation, we set $\tau = 0.5$, which provides a balanced trade-off between boundary sensitivity and segment continuity. Since the model only performs a single forward classification over $L - 1$ adjacent pairs, inference complexity grows approximately linearly with the number of sentences, and batch processing plus GPU acceleration further improves throughput, yielding substantial efficiency advantages over generative segmentation.

3.3. Generative-Alignment Query Expansion

Short interrogative queries and long declarative documents often occupy different regions of the embedding space, which weakens retrieval alignment in long-document RAG. A representative solution is Hypothetical Document Embeddings (HyDE), which generates a document-like expansion from the original query and uses it as a retrieval anchor. However, directly replacing the original query representation with the generated document embedding may introduce semantic drift when the generated content is imperfect. To address this issue, we propose Generative-Alignment Query Expansion (GAQE), which follows a two-step design: it first generates a context-rich hypothetical document to reduce the query-document modality gap, and then fuses the generated representation with the original query embedding through a residual weighting mechanism. This design preserves the semantic richness of document-form expansion while retaining the directional constraint of the original query, as illustrated in Figure 5.

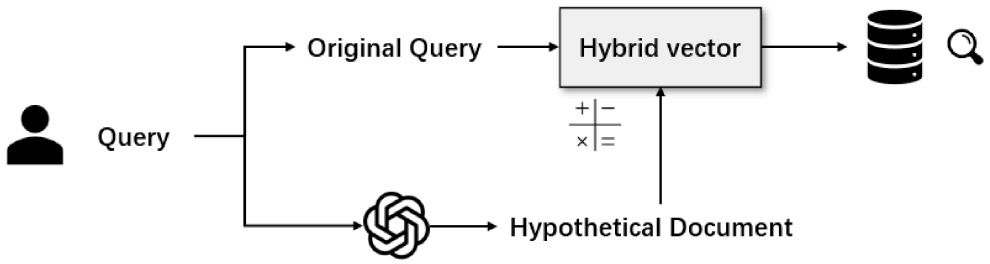


Figure 5. Workflow of generative-alignment query expansion.

3.3.1. Hypothetical Document Generation

Given an original query q, an LLM is first used to generate a hypothetical document $\hat{d}$ that resembles the style and density of relevant source passages. Instead of producing a conversational answer, the generator is guided by a structured prompt (Table 1) to output a high-density declarative paragraph containing contextual entities and domain-specific

terminology. The purpose of this step is to move the query representation closer to the document manifold in embedding space, thereby improving the retriever's ability to match semantically relevant evidence even when lexical overlap is limited.

Table 1. System Prompt for Hypothetical Document Generation.

Prompt: Hypothetical Document Generation	
Task	Generate a hypothetical document that contains high-density information to answer the query below.
Style Constraints	1. Declarative: Use objective, factual statements only. 2. Specific: Prioritize domain-specific terminology over general descriptions. 3. Concise: Do not include introductory or concluding fillers.
Negative Constraints	Do not explicitly reference "the user" or "the query". Avoid speculative language (e.g., "It is possible...").
Input	{user_query}

Formally, the generated hypothetical document is defined as:

$$\hat{d} = \text{LLM}(q, \text{Prompt}_{structure}) \tag{4}$$

In this design, the hypothetical document does not serve as the final answer. Instead, it acts as a retrieval-oriented semantic bridge between the short question form and the long declarative form of evidence passages.

3.3.2. Residual Weighted Fusion

Although hypothetical-document generation can improve semantic recall, directly retrieving with the embedding of $\hat{d}$ may over-amplify generation bias or hallucinated details. GAQE therefore introduces a residual weighted fusion strategy in which the original query embedding is retained as a semantic anchor.

Let $\text{Enc}(\cdot)$ denote the retrieval encoder. We first compute the embeddings of the original query and the hypothetical document:

$$v_q = \text{Enc}(q),\ v_{\hat{d}} = \text{Enc}\left(\hat{d}\right) \tag{5}$$

$$v_{hybrid} = \alpha \cdot v_q + (1 - \alpha) \cdot v_{\hat{d}} \tag{6}$$

where α is a balancing hyperparameter. In our implementation, we set $\alpha = 0.3$, so that the hypothetical document embedding provides the dominant semantic signal while the original query embedding is retained as a residual constraint anchor. This improves semantic recall while reducing the risk of excessive drift introduced by generated expansions.

3.4. Information Density Truncation (IDT)

To resolve signal-to-noise imbalance caused by static Top-K truncation under varying query scenarios, we propose a training-free adaptive truncation algorithm, IDT. Instead of using a fixed K, IDT dynamically determines the optimal boundary by maximizing the information density of the selected context window.

3.4.1. Information Density Function

Assume the retriever returns a ranked list of candidate passages $\mathcal{C}_{\text{ret}} = \{c_1, \ldots, c_N\}$, sorted in descending order by relevance scores $S = \{s_1, \ldots, s_N\}$. We treat the top-k passages as a context window and regard this window as a single information unit. To quantify the

effective information content of the window, we first define the cumulative relevance $R(k)$ as the sum of passage scores within the window:

$$R(k) = \sum_{i=1}^{k} s_i \tag{7}$$

Maximizing $R(k)$ alone would encourage overly long contexts, diluting core evidence and increasing generator cost. Therefore, we introduce a nonlinear length penalty k^{γ} and define the information density $D(k)$:

$$D(k) = \frac{R(k)}{k^{\gamma}} = \frac{\sum_{i=1}^{k} s_i}{k^{\gamma}} \tag{8}$$

where $\gamma \in [0, 1]$ is a global length-penalty factor. In our implementation, we set $\gamma = 0.8$, which provides a good balance between preserving useful evidence and suppressing low-value long-tail passages.

3.4.2. Dynamic Truncation Strategy

$D(k)$ essentially reflects the average signal-to-noise ratio of the current context window. As k increases (i.e., more passages are included), the cumulative relevance $R(k)$ grows monotonically. However, because the marginal relevance of lower-ranked passages s_k gradually decreases and the denominator k^{γ} continues to increase, the overall density $D(k)$ typically exhibits a rise-then-fall pattern. Our goal is to identify the globally optimal truncation point k^* that preserves high information content while minimizing redundant noise. This objective is formalized as:

$$k^* = argmax_{1 \le k \le N} D(k) \tag{9}$$

The detailed procedure is given in Algorithm 1. Finally, the system retains only the top k^* passages, $\mathcal{C}^* = \{c_1, \ldots, c_{k^*}\}$, as the input to the generator. This strategy requires no additional model training and achieves adaptive selection via a single linear scan, effectively avoiding noise injection in sparse settings and premature information truncation in dense settings.

Algorithm 1. Information Density Truncation.

Input	Ranked chunks $\mathcal{C}_{ret} = \{c_1, \ldots, c_N\}$ with scores $S = \{s_1, \ldots, s_N\}$ (sorted in descending order); length penalty γ
Output	Selected context $\mathcal{C}^* = \{c_1, \ldots, c_{k^*}\}$
1	Initialize $R \leftarrow 0$, $D_{\max} \leftarrow -\infty$, $k^* \leftarrow 1$
2	for $k = 1 \ldots N$ do
3	$R \leftarrow R + s_k$
4	$D(k) \leftarrow R/k^{\gamma}$
5	if $D(k) > D_{\max}$ *then*
6	$D_{\max} \leftarrow D(k)$
7	$k^* \leftarrow k$
8	end if
9	end for
10	Return $\mathcal{C}^* = \{c_1, \ldots, c_{k^*}\}$

Here, $D(k)$ denotes the information density evaluated at truncation position k. During the scan, the algorithm computes $D(k)$ sequentially for each candidate cutoff and only

retains the current best value $D_{\max}$ and its corresponding position k^*, rather than storing all density values.

4. Experiments

This section evaluates SES-RAG from several complementary perspectives. We first introduce the datasets, evaluation metrics, baselines, and implementation details. We then report the overall benchmark results on three long-document QA datasets, followed by segmentation-quality analysis and module ablation. Finally, we examine retrieval robustness across different retrievers and evaluate the efficiency of the segmentation stage.

4.1. Datasets and Evaluation Metrics

To assess the effectiveness of SES-RAG across different long-document QA settings, we evaluate the framework on three widely used benchmarks: NarrativeQA [20], QuALITY [21], and QASPER [22]. NarrativeQA focuses on abstractive question answering over long narrative documents, such as books and movie scripts, and is suitable for evaluating long-range evidence aggregation and generation quality. QuALITY is a multiple-choice reading comprehension benchmark with long documents and strong distractors, requiring precise retrieval and multi-paragraph reasoning. QASPER evaluates question answering over scientific papers, testing both factual evidence localization and reasoning under technical terminology.

For NarrativeQA, we report ROUGE, BLEU-1, BLEU-4, and METEOR. For QuALITY, we report Accuracy. For QASPER, we report F1-Match. All reported metrics are percentages and are rounded to two decimal places for consistency.

4.2. Baselines and Implementation Details

To provide a more rigorous comparison than a single minimal baseline, we evaluate SES-RAG against four representative retrieval-augmented generation settings. Naive RAG serves as the basic reference pipeline, using fixed-window chunking, dense retrieval with the original query, and fixed Top-K context truncation. HyDE-RAG augments retrieval with a hypothetical document generated from the query, following the idea of document-form query expansion. Hybrid-RAG combines sparse and dense retrieval signals to improve recall under lexical mismatch. CRAG is included as a stronger corrective retrieval baseline to represent more advanced retrieval enhancement beyond one-shot retrieval [23]. For fair comparison, all frameworks use the same generator, GPT-4o-mini, and the same document corpus within each dataset.

In the Naive RAG setting, documents are segmented using fixed-window chunking with a window size of 200 tokens and a stride of 200. Retrieval is performed using the original query embedding, and a fixed Top-K = 5 truncation strategy is used as the default context selection policy. Unless otherwise stated, this setting is also used as the shared reference configuration for the ablation experiments.

All training and inference were conducted on a single NVIDIA RTX 4090D (24 GB; NVIDIA Corporation, Santa Clara, CA, USA) server. The software stack uses Ubuntu 22.04, PyTorch 2.4.1, and CUDA 12.1. Table 2 lists details. For reproducibility, all compared frameworks use the same corpus split, generator, and evaluation protocol within each dataset.

SES-RAG introduces three main hyperparameters: the segmentation threshold τ, the fusion weight α, and the IDT length-penalty factor γ. We calibrated them through lightweight pilot experiments over limited candidate sets rather than exhaustive search. Specifically, we evaluated $\tau \in \{0.4, 0.5, 0.6\}$, $\alpha \in \{0.1, 0.3, 0.5, 0.7\}$, and $\gamma \in \{0.6, 0.8, 1.0\}$. The final values were selected according to overall stability and trade-

off behavior: $\tau = 0.5$ balanced boundary sensitivity and segment continuity, $\alpha = 0.3$ preserved the original query as an anchor while benefiting from document-style semantic expansion, and $\gamma = 0.8$ balanced useful evidence retention against long-tail noise suppression. After calibration, these values were fixed across all benchmarks to avoid dataset-specific overfitting and to ensure a fair cross-benchmark comparison.

Table 2. Experimental Configuration and Hardware Environment.

Configuration	Details
Docker Image	PyTorch 2.4.1 • Python 3.12 • Ubuntu 22.04 • CUDA 12.1
GPU	RTX 4090 D (24 GB) × 1
CPU	16 vCPU Intel Xeon Platinum 8481C
Memory	80 GB

4.3. Overall Benchmark Results

Table 3 presents the overall end-to-end benchmark results used for cross-method comparison. It summarizes the primary performance of Naive RAG, HyDE-RAG, Hybrid-RAG, CRAG, and SES-RAG on NarrativeQA, QuALITY, and QASPER under a common experimental configuration. Additional analyses from other perspectives, including module ablation, retriever robustness, and efficiency, are reported separately in Sections 4.4–4.6. Overall, SES-RAG achieves the best performance across all three benchmarks, indicating that lightweight cross-stage optimization over semantic segmentation, query alignment, and adaptive context selection is more effective than relying only on stronger retrieval heuristics or corrective retrieval alone.

Table 3. Main end-to-end benchmark results on NarrativeQA, QuALITY, and QASPER.

Dataset/Metric	Naive RAG	HyDE-RAG	Hybrid-RAG	CRAG	SES-RAG
NarrativeQA					
ROUGE	10.61	10.83	11.54	12.79	**15.71**
BLEU-1	9.26	9.29	9.37	9.54	**9.61**
BLEU-4	2.52	2.63	3.92	5.03	**6.16**
METEOR	11.62	11.67	13.21	14.29	**18.71**
QuALITY					
Accuracy	70.67	71.00	72.33	72.33	**74.33**
QASPER					
F1-Match	37.23	37.81	38.57	39.84	**41.06**

On NarrativeQA, SES-RAG achieves the strongest gains among all compared methods, reaching 15.71 ROUGE, 9.61 BLEU-1, 6.16 BLEU-4, and 18.71 METEOR. Compared with Naive RAG, the absolute improvements are +5.10 ROUGE, +0.35 BLEU-1, +3.64 BLEU-4, and +7.09 METEOR. Even compared with the strongest baseline, CRAG, SES-RAG still improves ROUGE by 2.92 points, BLEU-4 by 1.13 points, and METEOR by 4.42 points. These results suggest that SES-RAG is particularly effective for long-form generative QA, where semantically coherent chunk construction and adaptive context filtering are critical for preserving complete evidence and reducing noisy context accumulation. The relatively small change in BLEU-1, together with the much larger gains in BLEU-4 and METEOR, further indicates that the improvements mainly come from better higher-order semantic alignment rather than simple lexical overlap.

On QuALITY, SES-RAG achieves the highest accuracy of 74.33, outperforming Naive RAG by 3.66 points and exceeding both Hybrid-RAG and CRAG by 2.00 points. Since QuALITY often requires reasoning over long documents with strong distractors, this result shows that SES-RAG can better preserve and assemble the evidence needed for multi-paragraph reading comprehension. The limited gain of HyDE-RAG over Naive RAG also

suggests that query expansion alone is not sufficient in this setting; instead, more stable improvements require joint optimization of retrieval-unit quality, retrieval alignment, and context selection.

On QASPER, SES-RAG obtains the best F1-Match score of 41.06, improving over Naive RAG by 3.83 points and over CRAG by 1.22 points. This result demonstrates that the proposed framework remains effective in scientific QA scenarios, where evidence is often distributed across technically dense and structurally complex documents. The improvement suggests that SES-RAG not only enhances retrieval relevance, but also improves the precision of the final generator input by reducing low-value long-tail passages.

Overall, the results across all three benchmarks consistently support the effectiveness of SES-RAG. While HyDE-RAG, Hybrid-RAG, and CRAG each improve the baseline to different extents, SES-RAG delivers the strongest and most stable gains, indicating that long-document RAG benefits more from coordinated improvements across multiple stages of the pipeline than from isolated enhancement at a single stage.

4.4. Segmentation Quality and Ablation Study

We analyze SES-RAG from two perspectives: micro-level segmentation quality and macro-level module contributions.

4.4.1. Segmentation Quality Evaluation

To examine concerns that fixed-window splitting harms semantic coherence, we design a Semantic Completeness Score (SCS) experiment. We randomly sample 10 long documents from WikiText and segment them using: fixed-window splitting (200 window/50 overlap), recursive character splitting (an industry baseline), and our SES-Segmenter. We then use GPT-4o as a judge to rate each chunk's semantic self-containment from 1 to 5 (5 = fully self-consistent).

As shown in Figure 6, fixed-window splitting yields an average SCS of 1.86 (std 0.42) due to mechanical truncation, often producing unclear references or broken logic. Recursive splitting improves to 2.98, but tends to force unrelated topics together to meet length targets, diluting information. In contrast, SES-Segmenter achieves the best score of 4.02 with the smallest std (0.26). The significant advantage confirms that Module A provides more stable boundary recognition, ensuring each retrieval unit is a semantically complete "information atom" for downstream retrieval.

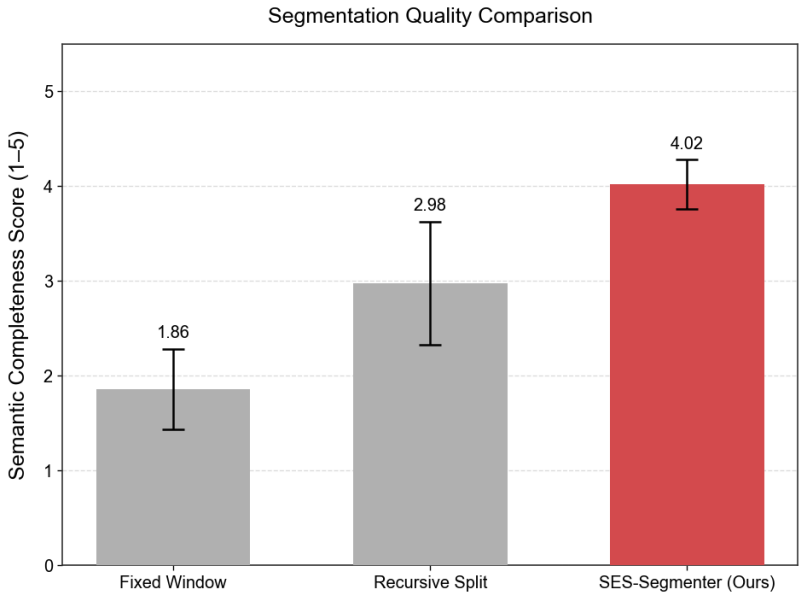


Figure 6. Segmentation Quality Comparison.

4.4.2. Ablation Study

Table 4 reports ablations on NarrativeQA by removing or replacing each module to quantify independent contributions.

Table 4. Ablation Study of SES-RAG Modules on NarrativeQA.

Model Configuration	ROUGE	BLEU-1	BLEU-4	METEOR
Naive RAG	10.61	9.26	2.52	11.62
Naive RAG + Module A	14.02	9.40	4.07	16.47
Naive RAG + Module B	11.32	9.41	4.15	11.73
Naive RAG + Module C	13.04	9.59	3.80	13.35
Naive RAG + GPT-4o Mini Seg.	14.48	9.51	4.59	16.51
SES-RAG	**15.71**	**9.61**	**6.16**	**18.71**

When removing semantic segmentation and reverting to fixed windows, ROUGE and METEOR drop most, indicating that even advanced retrieval/selection cannot recover coherence if the underlying units are fragmented. Module A is the foundation.

Removing GAQE causes BLEU-4 to drop by 1.63, suggesting that without modality bridging, key evidence is missed, harming precision in generated answers; Removing adaptive selection reduces overall generation quality, especially METEOR and BLEU-4, indicating that IDT helps improve the signal-to-noise ratio by filtering tail noise.

Together, the three modules specialize in preserving integrity, improving recall, and denoising context, and the combined system outperforms any single-module enhancement, validating the systematic design of SES-RAG.

4.5. Retrieval Robustness Across Retrievers

To validate generalization under different retrieval environments, we evaluate SES-RAG with three retrievers: OpenAI Embedding (strong dense retrieval), SBERT (standard dense retrieval), and BM25 (sparse retrieval).

As shown in Table 5, SES-RAG consistently outperforms Naive RAG under all retrievers. Notably, the relative gain is largest with BM25 (ROUGE +9.90), strongly supporting Module B's role: by generating keyword-rich hypothetical documents, GAQE effectively supplements sparse retrievers with missing semantic cues, providing a "semantic smoothing" effect. This suggests SES-RAG is robust even in constrained environments where expensive dense retrievers cannot be deployed.

Table 5. Retrieval Robustness of SES-RAG across Different Retrievers on NarrativeQA.

Retriever	Framework	ROUGE	BLEU-1	BLEU-4	METEOR
OpenAI Embedding	Naive RAG	10.61	9.26	2.52	11.62
OpenAI Embedding	**SES-RAG**	**15.71**	**9.61**	**6.16**	**18.71**
SBERT	Naive RAG	7.72	8.39	3.92	7.60
SBERT	SES-RAG	15.21	9.01	4.55	16.75
BM25	Naive RAG	4.97	6.99	5.15	4.55
BM25	SES-RAG	14.87	7.26	5.56	14.87

4.6. Efficiency Evaluation

To assess deployment feasibility, we compare inference overhead in the segmentation stage. Under the same hardware and document set, we segment 10 long documents using: (1) GPT-4o-mini generative segmentation; and (2) ModernBERT discriminative segmentation (Module A).

Figure 7 shows that generative segmentation provides usable quality but incurs high average latency (1015.2 s/document) and external API cost. In contrast, Module A performs discriminative inference on adjacent sentence pairs and benefits from batching and Flash Attention optimizations, reducing average processing time to 21.8 s/document—about a 46× speedup. This demonstrates SES-RAG's suitability for large-scale offline indexing and cost-sensitive online systems.

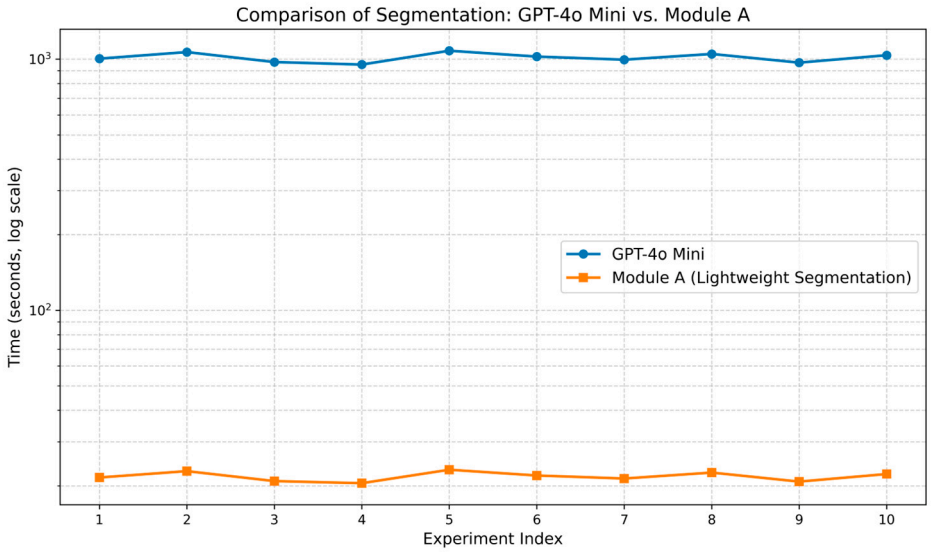


Figure 7. Comparison of Segmentation: GPT-4o Mini vs. Module A.

GPT-4o-mini segmentation is autoregressive and requires external API-based text generation over long contexts, which introduces both token-generation delay and service overhead. By contrast, Module A performs discriminative boundary classification over adjacent sentence pairs with a single forward-encoding workflow, and further benefits from GPU batching and Flash Attention acceleration. Therefore, the comparison reflects the efficiency difference between generative segmentation and lightweight discriminative segmentation under the same document set and hardware conditions, rather than a minor implementation-level optimization.

5. Conclusions and Future Work

This paper presents SES-RAG, a semantics-enhanced and adaptive optimization framework for long-document retrieval-augmented generation. The framework is designed to address three recurring bottlenecks in long-document RAG systems: semantic fragmentation during document chunking, query–document mismatch during retrieval, and signal-to-noise imbalance during context selection. To address these issues in a lightweight and modular manner, SES-RAG introduces three complementary components: a ModernBERT-based semantic segmenter for semantically coherent chunk construction, GAQE for query expansion and alignment through residual fusion with a hypothetical document, and IDT for adaptive context truncation based on retrieval-score-derived information density.

Experimental results on NarrativeQA, QuALITY, and QASPER show that SES-RAG consistently achieves the best overall performance among the compared methods, including

Naive RAG, HyDE-RAG, Hybrid-RAG, and CRAG. The gains are especially evident on long-form generative question answering, where preserving coherent retrieval units and filtering noisy context are particularly important. The ablation study further shows that the three modules make complementary contributions, while the cross-retriever experiments demonstrate that the effectiveness of SES-RAG is not restricted to a single retrieval backbone. These findings suggest that improving long-document RAG requires not only stronger retrieval, but also better coordination across segmentation, query alignment, and adaptive context selection.

Compared with heavier corrective or iterative RAG frameworks, SES-RAG provides a practical alternative that improves multiple stages of the pipeline without introducing excessive architectural complexity. This makes it suitable for applications that require a balance between answer quality, efficiency, and ease of deployment.

Future work can be pursued in several directions:

1. First, multimodal adaptation is a promising extension. The current framework focuses on plain-text long documents, while many real-world documents contain figures, tables, layout cues, and other visual signals that may help semantic segmentation and evidence selection. Incorporating such multimodal information could further improve the framework in document QA scenarios involving PDFs, reports, and scientific articles.
2. Second, structure-aware retrieval enhancement could be explored to strengthen semantic alignment beyond linear textual expansion. For example, integrating knowledge graphs or graph-based representations may help model cross-sentence entity relations and improve retrieval over structurally complex documents.
3. Third, the current framework uses fixed hyperparameters selected through small-scale validation. A more flexible direction would be to design query-aware adaptive parameterization, allowing the system to dynamically adjust expansion and truncation behavior according to query complexity, retrieval confidence, or document structure.

In summary, SES-RAG demonstrates that lightweight cross-stage optimization is an effective way to improve long-document RAG. By jointly enhancing semantic segmentation, query alignment, and adaptive context selection within a standard pipeline, the proposed framework offers a practical and extensible solution for building more accurate and robust retrieval-augmented generation systems.

Author Contributions: Conceptualization, C.G. and Q.Z.; methodology, H.H.; software, H.H.; validation, H.H., C.G. and Q.Z.; formal analysis, H.H.; investigation, H.H.; resources, C.G. and Q.Z.; data curation, H.H.; writing—original draft preparation, H.H.; writing—review and editing, H.H., C.G. and Q.Z.; visualization, H.H.; supervision, C.G. and Q.Z.; project administration, C.G. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study used publicly available datasets, including NarrativeQA, QuALITY, and QASPER. The code and related materials will be released publicly after the organization process is completed. https://github.com/HereAfterUs/SES-RAG (accessed on 20 April 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
2. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
3. Pan, Z.; Wu, Q.; Jiang, H.; Xia, M.; Luo, X.; Zhang, J.; Lin, Q.; Rühle, V.; Yang, Y.; Lin, C.-Y.; et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 963–981.
4. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In Proceedings of the Twelfth International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
5. Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 2526–2547.
6. Zhao, J.; Ji, Z.; Fan, Z.; Wang, H.; Niu, S.; Tang, B.; Xiong, F.; Li, Z. Moc: Mixtures of text chunking learners for retrieval-augmented generation system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 5172–5189.
7. Duarte, A.V.; Marques, J.D.; Graça, M.; Freire, M.; Li, L.; Oliveira, A.L. Lumberchunker: Long-form narrative document segmentation. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 6473–6486.
8. Zhong, Z.; Liu, H.; Cui, X.; Zhang, X.; Qin, Z. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics, 2025, Abu Dhabi, UAE, 19–24 January 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 5756–5774.
9. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 2318–2335.
10. Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1762–1777.
11. Su, W.; Tang, Y.; Ai, Q.; Wu, Z.; Liu, Y. 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)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 12991–13013.
12. Zhang, Z.; Fang, M.; Chen, L. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 6963–6975.
13. Ye, F.; Li, S.; Zhang, Y.; Chen, L. R2AG: Incorporating retrieval information into retrieval augmented generation. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 11584–11596.
14. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173. [CrossRef]
15. Yu, Y.; Ping, W.; Liu, Z.; Wang, B.; You, J.; Zhang, C.; Shoeybi, M.; Catanzaro, B. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 121156–121184.
16. Jiang, H.; Wu, Q.; Luo, X.; Li, D.; Lin, C.-Y.; Yang, Y.; Qiu, L. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 1658–1677.
17. Frohmann, M.; Sterner, I.; Vulić, I.; Minixhofer, B.; Schedl, M. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 11908–11941.
18. Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **2024**, *568*, 127063. [CrossRef]
19. Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16344–16359.

20. Kočiský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K.M.; Melis, G.; Grefenstette, E. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 317–328. [CrossRef]
21. Pang, R.Y.; Parrish, A.; Joshi, N.; Nangia, N.; Phang, J.; Chen, A.; Padmakumar, V.; Ma, J.; Thompson, J.; He, H. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022.
22. Dasigi, P.; Lo, K.; Beltagy, I.; Cohan, A.; Smith, N.A.; Gardner, M. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021.
23. Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; Ling, Z.-H. Corrective retrieval augmented generation. In *ICLR 2025 Conference Withdrawn Submission*; 2024. Available online: https://openreview.net/forum?id=JnWJbrnaUE (accessed on 20 April 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.