GGatrieval: Fine-grained Grounded Alignment Retrieval for Verifiable Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) systems bolster large language models (LLMs) by integrating retrieval mechanisms to overcome limitations in knowledge scope. However, traditional retrieval mechanisms, which predominantly operate at the sentence level, often fail to capture complete semantics at finer syntactic constituent granularities, degrading generation quality. To address this, we propose GGatrieval (Fine-**g**rained **G**rounded **A**lignment Re**trieval** for Verifiable Generation), a novel framework that enhances retrieval by targeting syntactic constituent interactions. Specifically, drawing inspiration from human cognitive processes, GGatrieval introduces a document selection criterion and assigns categorical labels via a Fine-grained Grounded Alignment strategy. These labels enable document reranking and drive a Semantic Compensation Query Augmentation strategy, yielding enriched queries that retrieve documents tightly aligned with the original query. Experiments on the ALCE benchmark and the extended Natural Questions datasets demonstrate GGatrieval's superior performance over established baselines, with ablation studies validating the effectiveness of our selection criterion and classification methods.

1 Introduction

Retrieval-Augmented Generation (RAG) systems integrate large language models (LLMs) with targeted retrieval mechanisms to address knowledge coverage limitations of generative models (Lewis et al., 2020). By retrieving relevant external knowledge, RAG improve output accuracy (Khandelwal et al., 2019; Min et al., 2020), mitigates LLM hallucinations (Cheng et al., 2024), and incorporates current real-world information (Gupta et al., 2024), often without additional model training (Izacard et al., 2023a).

The retrieval mechanism in RAG systems comprises three pivotal stages: Pre-retrieval, Retrieval, and Post-retrieval. In the Pre-retrieval stage, indexing leverages methods like graphs, product quantization (PQ) (Liu et al., 2023a), and locality-sensitive hashing (LSH) (Datar et al., 2004), employing approximate nearest neighbor search (ANNS) (Arya et al., 1998) for efficiency. Query manipulation, including query expansion, reformulation, and prompt-based rewriting (Izacard and Grave, 2021; Wang et al., 2023; Chan et al., 2024; Zheng et al., 2023), refines queries to address ambiguities, significantly boosting retrieval accuracy. The Retrieval stage employs search and ranking techniques such as CRAG, IRCOT, and FLARE (Yan et al., 2024a; Trivedi et al., 2023; Jiang et al., 2023), optimizing document relevance via few-shot learning and confidence-based strategies. Retrieval strategies—basic, iterative, recursive, conditional, and adaptive (Shao et al., 2023; Kang et al., 2023; Yue et al., 2024; Asai et al., 2024)—tailor the process to specific tasks, enabling dynamic, context-

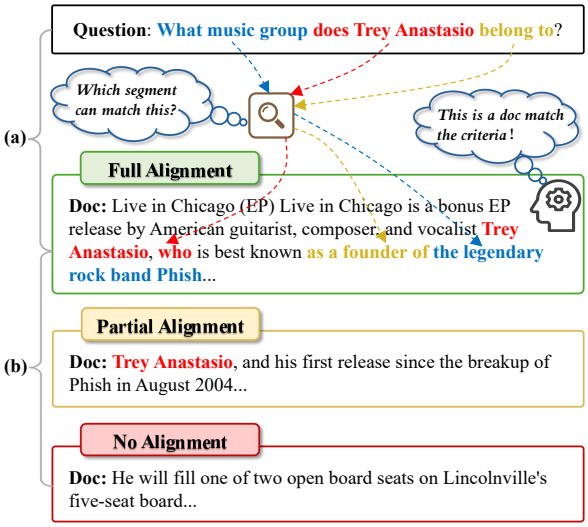


Figure 1: Document Selection Criteria and Document Taxonomy. (a) refers to human cognitive process for acquiring standard documents. (b) refers to examples of labels for different document categories.

sensitive retrieval. In the Post-retrieval stage, reranking, using unsupervised and supervised methods alongside data augmentation (Ram et al., 2023; Ma et al., 2024; Sun et al., 2023), prioritizes pertinent documents, while filtering techniques like Self-RAG and RECOMP (Asai et al., 2024; Xu et al.) eliminate irrelevant content, enhancing output quality. Collectively, these stages ensure RAG systems retrieve and refine information effectively, improving relevance and accuracy in knowledge-intensive tasks.

Limitation. However, conventional retrieval mechanisms typically operate at the sentence level, leading to semantic incompleteness at the syntactic constituent granularity. This deficiency means retrieved documents may lack the semantic information needed fully address queries, ultimately limiting the generation quality of RAG systems.

Our approach. The meaning of complex expressions derives from their fundamental components (Drozdov et al., 2022). Syntactic parsing particularly crucial for sentence understanding (Lesmo and Lombardo, 1992). Consequently, a human cognitive process for selecting retrieval documents can be summarized as follows, also as shown in Figure 1(a): (1) Decompose the user query into basic syntactic constituents; (2) Identify continuous textual segments in candidate documents that semantically match these constituents; (3) Determine that a candidate document fully supports query-answer generation if it contains a segment aligning with all query constituents.

Inspired by these insights, we propose GGatrieval (Fine-**g**rained **G**rounded **A**lignment Re**trieval** for Verifiable Generation), a framework that enhances retrieval by aligning information with the user query at the syntactic constituent level and validates it through Verifiable Generation (Gao et al., 2023). Specifically, we introduce a novel document selection criterion, which assesses whether a continuous textual segment within a retrieval document semantically aligns with every syntactic constituent of the query. Based on this criterion, we define category labels for retrieval documents, as shown in Figure 1(b), derived through a Fine-grained Grounded Alignment (**FGA**) strategy. These labels enable document re-ranking and underpin a Semantic Compensation Query Augmentation (**SCQA**) strategy. By performing semantic compensation at the syntactic constituent granularity, SCQA generates diverse, semantically rich augmented queries, retrieving documents highly aligned with the original query.

Experiments on the ALCE benchmark (Gao et al., 2023) and extended Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) demonstrate that GGatrieval outperforms mainstream baselines. Ablation studies and analysis validate the effectiveness of our document selection criterion and the role of document labels. Notably, on the ELI5 dataset, GGatrieval improves Claim F1 by 22% and Citation F1 by 28%.

Contributions. Our key contributions are summarized below.

- We propose a novel document selection criterion, enabling precise document classification and validating its effectiveness.
- We propose a FGA strategy, enhancing document verifiability within RAG systems, thus improving the credibility of generated outcomes.
- We propose a SCQA strategy to bridge the semantic gap between queries and target documents, thereby enhancing retrieval document quality.
- Extensive experiments across various datasets and retrieval optimization baselines demonstrate the superior performance of our approach compared to existing methods.

2 Related Work

2.1 Verifiable Generation

Verifiable generation refers to producing text that can be independently traced and validated through explicit citations. Current methodologies fall into two main categories. The first involves directly embedding citations during text generation by leveraging the inherent capabilities of language models. For instance, Weller et al. (2024) prompt LLMs with citation cues (e.g., "according to Wikipedia"), while Lee et al. (2023) systematically evaluate and provide feedback on text quality, guiding models toward improved verifiability. The second, retrieval-based approach emphasizes accurate citations by incorporating external sources, such as webpages or documents. Notably, WebGPT (Nakano et al., 2021) and LaMDA (Thoppilan et al., 2022) construct large-scale training datasets from web and Wikipedia resources, enabling citation-rich outputs. Additionally, Li et al. (2024) iteratively refine citation quality by aligning retrieved content with

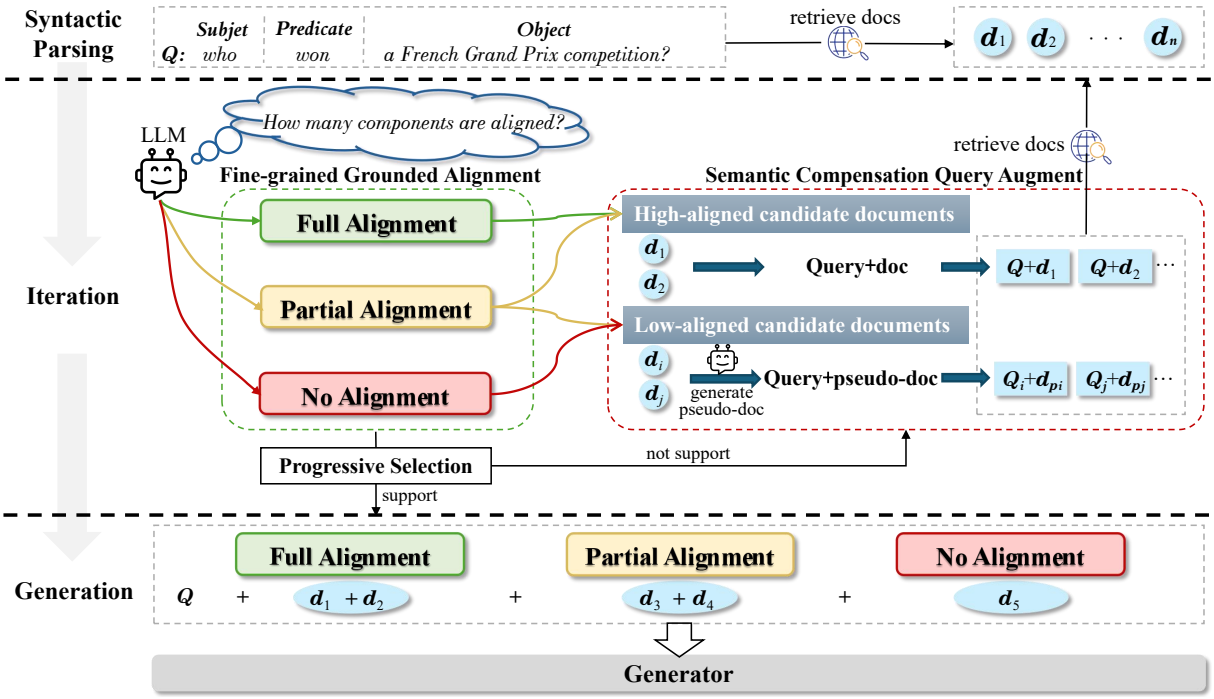


Figure 2: Overview of GGatrieval at inference. Our approach defines categorical labels {Full Alignment, Partial Alignment, No Alignment} for documents (Section 3.1) and employs a FGA strategy (Section 3.2) to assign appropriate labels to each document. Furthermore, based on these categorical labels, we use a SCQA (Section 3.3) strategy to retrieve high-quality documents.

generated responses. Our method adopts the latter retrieval-oriented paradigm, facilitating rigorous comparison of retrieval mechanisms within RAG frameworks.

2.2 Retrieval mechanisms

The retrieval mechanism in RAG systems comprises three critical stages: Pre-retrieval, Retrieval, and Post-retrieval, each enhancing precision and contextual relevance. In the Pre-retrieval stage, efficient indexing utilizes graphs, product quantization (PQ) (Liu et al., 2023a), and locality-sensitive hashing (LSH) (Datar et al., 2004), supported by approximate nearest neighbor search (ANNS) (Arya et al., 1998). Query manipulation techniques, including query expansion, reformulation, and prompt-based rewriting (Izacard and Grave, 2021; Wang et al., 2023; Chan et al., 2024; Zheng et al., 2023), refine input queries to address ambiguity and enhance retrieval accuracy. The Retrieval stage incorporates search and ranking approaches such as Atlas, AAR, IRCOT, and FLARE (Izacard et al., 2023b; Yu et al., 2023; Trivedi et al., 2023; Jiang et al., 2023), utilizing strategies like few-shot learning and dynamic adaptation (Shao et al., 2023; Kang et al., 2023; Yue et al., 2024; Asai et al., 2024). Post-retrieval, re-ranking and filtering methods—including Self-RAG, RECOMP, and CRAG (Ram et al., 2023; Ma et al., 2024; Asai et al., 2024; Xu et al.; Yan et al., 2024a)—further refine results, ensuring RAG systems deliver highly relevant and accurate information for knowledge-intensive applications.

3 Methods

Figure 2 presents an overview of GGatrieval during inference. This method enhances document reliability through an optimized retrieval mechanism, thereby improving overall system performance. Specifically, upon receiving a user query, the system initially retrieves candidate documents using a conventional retriever. It then applies the FGA strategy (Section 3.2) to assign category labels (Section 3.1) to each document. Subsequently, the system re-ranks these candidate documents according to their matching degrees and relevance scores, selecting the final retrieval documents via a progressive selection algorithm, which are then verified. It then applies the FGA strategy (Section 3.2) to assign category labels (Section 3.1) to each document. Otherwise, a SCQA strategy (Section 3.3) is added to refine the query iteratively until a predefined iteration limit is reached. For more details regarding the specific algorithmic procedure,

please refer to Appendix C.

3.1 Selection Criterion and Document Taxonomy

Inspired by human cognitive processes for document selection, as shown in Figure 1(a), we propose a novel document selection criterion: whether a document contains a continuous text segment semantically aligning with all syntactic constituents of the query. Based on this, we define document labels into three types:

Full Alignment: A candidate document contains a continuous text segment semantically matches all syntactic constituents of the query.

Partial Alignment: A candidate document contains a continuous text semantically matches at least one syntactic constituent of the query, but does not match all syntactic constituents.

No Alignment: A candidate document contains no continuous text segment that semantically matches any syntactic constituent of the query.

Examples are shown in Figure 1(b). We use the FGA strategy (Section 3.2) to label retrieval documents. Subsequently, which are then re-ranked and filtered to support generation effectively.

3.2 Fine-grained Grounded Alignment

We propose a FGA strategy, which assigns specific category label to each retrieval document to represent the degree of semantic alignment between the retrieved document and the user query.

Query Syntactic Parsing: Given a user query Q, the system decomposes it into essential syntactic constituents such as subject, predicate, and object:

$$C = \{C_s, C_v, C_o, C_c, C_{attr}, C_{adv}, C_{supp}, C_{app}\} = LLM(I_Q, Q) \quad (1)$$

where I_Q denotes the instruction for syntactic parsing, and C is the set of extracted syntactic constituents from query Q. The components C_s,C_v,C_o,C_c,C_{attr},C_{adv},C_{supp} and C_{app} correspond to subject, predicate, object, predicative, attribute, adverbial, complement, and apposition, respectively. Leveraging LLM for shallow parsing enables comparable performance to traditional supervised methods without additional training or complex technical processing, improving the efficiency of system implementation.

Fine-grained Grounded Alignment: For each syntactic constituent of the query, the system analyzes candidate documents D to identify semantically matching fragments, yielding both an analytical process and the matched constituents. This process can be formulated as:

$$AnalysisResult = LLM(I_C^D, C_i, D) \quad (2)$$

where I_C^D prompts the LLM to analyze and match query syntactic constituents within document D. C_i denotes each syntactic constituent in C, $Analysisresult$ includes both the matching analysis and the corresponding results. Documents aligned with more syntactic constituents are more likely to support accurate answers. This process mirrors human strategies for document selection, demonstrating the potential of LLM-driven interaction with external corpora at the syntactic level.

Result Reflection: As emphasized by Liu et al. (2024), LLMs enhance performance through self-reflection capabilities. Therefore, we incorporate a reflection step to reassess the previous analysis as follows:

$$L(Q, D) = LLM(I_R, AnalysisResult) \quad (3)$$

Here, I_R instructs the LLM to reflect on the $AnalysisResult$, yielding a final list of matched syntactic constituents $L(Q, D)$ from the user query for document D.

Document Labeling: For document classification, we define the following symbols: let $|Q|$ be the number of syntactic constituents in the query Q, and $|L(Q, D)|$ the numberin $L(Q, D)$. Based on the document classification in Section 3.1, documents are labeled as follows: " Full Alignment" if $|L(Q, D)|$=$|Q|$, " Partial Alignment" if 0< $|L(Q, D)|$<$|Q|$, and " No Alignment" if $|L(Q, D)|$=0. These labels support subsequent document re-ranking and filtering, and also provide the semantic foundation for SCQA strategy (Section 3.3).

3.3 Semantic Compensation Query Augmentation

Dense retrievers excel at finding documents that are semantically related to the original query (Karpukhin et al., 2020; Lewis et al., 2020). Building on this, we propose a SCQA strategy, which generates augmented queries to retrieve highly semantically relevant documents:

Query Diversification: Full Alignment documents meet the target retrieval criterion and thus require

no further processing. However, for Partial Alignment or No Alignment documents, we generate synonymous queries based on syntactic constituents missing from $L(Q, D)$, as represented by:

$$Q' = \{LLM(I_s^{C_i}, C_i)\ for\ C_i\ in\ C \\ if\ C_i\ not\ in\ L(Q, D)\ else\ C_i\} \quad (4)$$

If a component C_i is not in $L(Q, D)$, $I_S^{C_i}$ prompts the LLM generates a synonymous description, otherwise, the original component is retained. Original and synonymous components jointly reconstruct the query, yielding diversified updated queriy Q'. This method leverages the diversity of $L(Q, D)$ to generate multiple queries that are semantically similar but differ in form.

Semantic Compensation Query Augmentation: Dense retrievers typically perform semantic similarity matching at the sentence level. Although query diversification enriches the original queries from various syntactic perspectives, there remain notable semantic differences at the level of syntactic constituent granularity between these queries and the target document. To bridge this semantic, we propose compensating semantic information at the syntactic constituent level, constructing enhanced queries that closely align with the target documents. The detailed implementation includes:

$$D_{\text{pseudo}} = LLM(I_{\text{pseudo}}, Q'), \quad if\ \frac{|L(Q, D)|}{|Q|} < \tau \quad (5)$$

$$Q'' = \begin{cases} Q + D, & if\ \frac{|L(Q,D)|}{|Q|} \geq \tau \\ Q' + D_{\text{pseudo}}\,, & if\ \frac{|L(Q,D)|}{|C|} < \tau \end{cases} \quad (6)$$

A document is considered high-aligned if $\frac{|L(Q,D)|}{|Q|} \geq \tau$, and Low-aligned otherwise. The threshold τ is user-defined and adjustable, controlling the intensity of semantic compensation and computational cost. For High-aligned documents D, we directly concatenate the query Q and the document D to form augmented queries. Given their semantic alignment, such concatenation enriches the query's semantic content. For Low-aligned documents, in order to compensate semantic information, we use I_{pseudo} to instruct the LLM to generate pseudo-documents D_{pseudo} aligned with every query syntactic constituent of the updated query Q', then we concatenated them to form augmented queries. This strategy dynamically updates query Q at each retrieval iteration, bridging semantic gaps between the query and target documents. Consequently, the system retrieves documents with high grounded alignment. Compared to LLatrieval (Li et al., 2024), GGatrieval reduces retrieval volume by 95% on the ASQA dataset and by 67% on the QAMPARI dataset, enhancing retrieval efficiency (Appendix B.4).

4 Experiment Settings

4.1 Datasets and Evaluation Metrics

In this study, experiments were conducted on the ALCE benchmark (Gao et al., 2023). Given that ALCE primarily focuses on multi-hop question answering, we expanded the LLatrieval baseline by incorporating a representative open-domain single-hop QA dataset—Natural Questions (NQ) (Kwiatkowski et al., 2019)—to achieve a comprehensive evaluation of GGatrieval. (1) ASQA (Stelmakh et al., 2022): An open-domain long-form QA dataset providing comprehensive and explanatory long answers to ambiguous factual questions. Answering questions from this dataset requires integrating multiple sources of information, resolving contextual ambiguities, and correlating various short answers; hence, ASQA is classified as a multi-hop dataset. (2) QAMPARI (Amouyal et al., 2023): A challenging open-domain QA benchmark specifically designed to handle multi-answer questions distributed across different paragraphs. (3) ELI5 (Fan et al., 2019): Developed by Facebook AI Research, designed to enhance AI models' capabilities in addressing complex explanatory questions and generating paragraph-level, multi-sentence answers. (4) NQ: The Natural Questions dataset, developed by Google Research, primarily used for evaluating machine reading comprehension tasks, emphasizing answer localization and extraction.

We evaluate the system's correctness and the verifiability of the documents using the ALCE framework proposed by Gao et al. (2023). For more details of datasets and evaluation metrics, please refer to Appendix A.1.

4.2 Baselines

To conduct a comprehensive evaluation, we selected seven representative baselines from the three stages of retrieval mechanisms, highlighting the advantages of GGatrieval in terms of document verifiability and overall system accuracy.

Dataset	ASQA				QAMPARI				ELI5				Overall	
	Correct	Citation			Correct	Citation			Correct	Citation			Correct	Citation F1
	EM-R	Rec	Prec	F1	F1	Rec	Prec	F1	Claim	Rec	Prec	F1		
BM25	36.57	36.79	38.98	37.86	10.21	18.95	19.67	19.3	11.65	41.28	41.4	23.47	19.48	26.88
BGE-E-large	51.57	53.63	55.95	54.73	12.06	26.42	27.48	26.93	-	-	-	-	-	-
CRAG	46.29	47.15	50.19	48.59	11.9	20.1	20.73	20.41	12.79	48.07	48.96	48.49	23.66	39.16
GGatrieval	**52.86**	**56.93**	**58.19**	**57.51**	**17.85**	**35.44**	**36.58**	**35.98**	**14.21**	**56.16**	**56.26**	**56.17**	**28.31**	**49.89**

Table 1: Comparison with Baselines in the Retrieval stage (USING gpt-3.5-turbo), The bolded numbers indicate the best performance.

Dataset	ASQA				QAMPARI				ELI5				Overall	
	Correct	Citation			Correct	Citation			Correct	Citation			Correct	Citation F1
	EM-R	Rec	Prec	F1	F1	Rec	Prec	F1	Claim	Rec	Prec	F1		
RankGPT	49.76	51.48	54.71	53.04	16.4	33.1	34.24	33.66	11.6	42.38	43.12	42.74	25.92	43.15
LLatrieval	50.8	53.54	55.75	54.58	16.86	34.09	34.9	34.46	11.62	43.47	44.85	43.88	26.43	44.31
GGatrieval	**52.86**	**56.93**	**58.19**	**57.51**	**17.85**	**35.44**	**36.58**	**35.98**	**14.21**	**56.16**	**56.26**	**56.17**	**28.31**	**49.89**

Table 2: Comparison with Baselines in the Post-retrieval stage (USING gpt-3.5-turbo).

Dataset	ASQA		QAMPARI	
	EM-R	Citation-F1	Correct-F1	Citation-F1
Query2Doc	50.44	53.04	16.76	33.26
MuGI	50.76	51.71	16.52	32.23
GGatrieval	**52.12**	**57.16**	**16.86**	**35.57**

Table 3: Comparison with baselines in the Pre-retrieval stage (USING gpt-3.5-turbo).

Dataset	NQ			
	Correct-F1	Rec	Prec	Citation-F1
BM25	20.08	31.28	33.61	32.41
BGE-E-large	26.06	31.5	34.95	33.13
CRAG	23.98	23.7	26.19	24.588
RankGPT	27.83	33.45	36.76	35.03
LLatrieval	27.3	32.64	36.03	34.25
GGatrieval	**28.68**	**34.51**	**37.58**	**35.98**

Table 4: Comparison with baselines on the NQ dataset (USING gpt-3.5-turbo).

Pre-retrieval stage: (1) MuGI (Zhang et al., 2024), which utilizes LLMs to generate multiple pseudo-reference documents combined with the original query to enhance sparse and dense retrieval effectiveness; and (2) Query2Doc (Wang et al., 2023), employing a few-shot prompting approach to generate pseudo-documents related to the query through LLMs, which are then appended to the original query to enhance expressiveness.

Retrieval stage: (1) BM25 (Robertson et al., 2009), a probabilistic model widely used in information retrieval for assessing the relevance between queries and documents; (2) BGE Large (Liu et al., 2023b), a general embedding model efficiently converting textual data into low-dimensional dense vectors, enabling effective semantic similarity computation and retrieval; and (3) CRAG (Yan et al., 2024b), introducing a lightweight retrieval evaluator for assessing the relevance and quality of retrieved documents given a specific query.

Post-retrieval stage: (1) RankGPT (Sun et al., 2023), which directly generates document rankings through language modeling; and (2) LLatrieval (Li et al., 2024), which improves retrieval quality via language model feedback, thereby supporting more accurate and verifiable generation. For details of baselines, please refer to Appendix A.2.

4.3 Implementation Details

We utilize the Verifiable Generation (Gao et al., 2023; Li et al., 2024) paradigm for both answer generation and evaluation. The APIs of OpenAI's " gpt-3.5-turbo" language model and the open-source " Meta-Llama3-8B-Instruct" model are used for GGatrieval, with the temperature set to 0 to minimize random variation. The threshold τ (Section 3.3) for document alignment categorizing is set to 0.66, determined through systematic experimentation and analysis. The number of supporting documents is set to 5, and the maximum number of iterations is 4. For ASQA, QAMPARI, and NQ datasets, the retrieval corpus is based on the Wikipedia dataset used in ALCE (Gao et al., 2023), with the dense embedding model BGE-large (Xiao et al., 2024) as the retriever. For the overall imple-

mentation details, please refer to Appendix A.3.

4.4 Main Results

Exp-1: Comparison with baselines in the Pre-retrieval stage. We evaluated GGatrieval against two query expansion baselines in the Pre-retrieval stage, MuGI and Query2Doc, on ASQA and QAMPARI datasets. To ensure a fair comparison, only the initial application of GGatrieval's SCQA strategy was employed. As shown in Table 3, GGatrieval outperformed the best baseline by 4.8% for Correct and 8.4% for Citation-F1 on ASQA, and by 0.5% for Correct and 7% for Citation-F1 on QAMPARI. Notably, larger improvements in Citation-F1 suggest GGatrieval retrieves more reliable documents, while baseline methods sometimes relied on lower-quality documents coincidentally producing correct answers.

Exp-2: Comparison with Baselines in the Retrieval stage. We assessed GGatrieval against conventional retrieval methods, BM25 and BGELarge, and a trained retriever, CRAG, using the ALCE benchmark. As shown in Table 1, GGatrieval surpassed the best baseline by 2.5% in Correct and 2.98% in Citation-F1 on ASQA. On QAMPARI, it achieved substantial improvements of 48% in Correct and 33.6% in Citation-F1. For the ELI5 dataset, GGatrieval improved Claim by 11.1% and Citation-F1 by 15.83%. These results demonstrate that the optimizations applied by GGatrieval in both the Pre-retrieval(SCQA) and Post-retrieval(FGA) stages are critical to its effectiveness.

Exp-3: Comparison with Baselines in the Post-retrieval stage. We compared GGatrieval with two baselines, RankGPT and LLatrieval, with results detailed in Table 2. On ASQA, GGatrieval improved Correct by 4.1% and Citation-F1 by 5.4%. On QAMPARI, it enhanced Correct by 5.9% and Citation-F1 by 4.4%. The most significant improvements occurred on ELI5, with a 22.3% increase in Claim and a 28% increase in Citation-F1. The ELI5 dataset exhibits greater performance improvements because its queries contain more redundant information, leading to a higher number of syntactic components and a greater diversity of queries. This increased query diversity enables GGatrieval to access more reliable documents. Detailed statistics are shown in Table 11 in Appendix B.4.

To evaluate the generality and robustness of our GGatrieval, we conducted experiments using the Meta-Llama3-8B-Instruct model. Please refer to Appendix B.1 for more details.

	ASQA		QAMPARI	
	EM-R	Cite	Correct-F1	Cite
Final Result	58.30	61.30	17.90	35.42
— Full Alignment	**46.79**	**53.43**	**13.43**	**28.52**
— Partial Alignment	48.38	53.13	15.04	33.65
— No Alignment	50.14	56.14	16.60	34.88

Table 5: Ablation study for excluding documents with different labels. " —" signifies that a specific category of document has been eliminated. The bolded numbers indicate the worst performance.

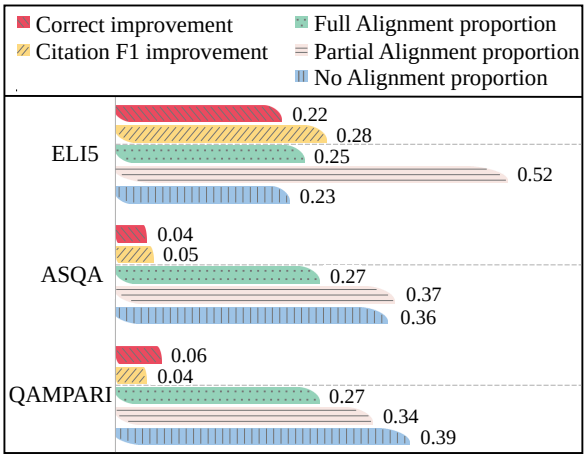


Figure 3: Cross-dataset analysis of label proportions. The bars above the dashed line represent system performance, while the bars below the dashed line indicate the proportion of documents with different labels.

Exp-4: Comparison with Baselines on the NQ Dataset To evaluate GGatrieval's effectiveness in single-hop question-answering tasks, we performed experiments on the NQ dataset, benchmarking it against multiple baselines. As shown in Table 4, GGatrieval consistently outperformed all five baselines from both the Retrieval and Post-retrieval stages. In particular, GGatrieval achieved a 3% improvement in Correct scores and a 2.7% increase in Citation-F1 over the strongest baseline. These results demonstrate the effectiveness of GGatrieval in both single-hop and multi-hop question answering tasks.

4.5 Ablation Study and Analysis

Exp-5: Ablation Study on Alignment Labels. We investigated the effect of alignment labels by systematically excluding documents labeled as "Full Alignment", "Partial Alignment", and "No Alignment" from the candidate retrieval set. Table 5 presents results for ASQA and QAMPARI datasets. To maintain comparability, remaining documents were re-ranked by alignment degree

and relevance, filling any gaps caused by exclusion. Excluding " Full Alignment" documents caused the most significant performance drop, followed by " Partial Alignment" , while " No Alignment" exclusions had minimal impact. This establishes a clear hierarchy of importance: " Full Alignment" > " Partial Alignment" > " No Alignment" , corroborated by additional analysis in Exp-10 (Appendix B.3).

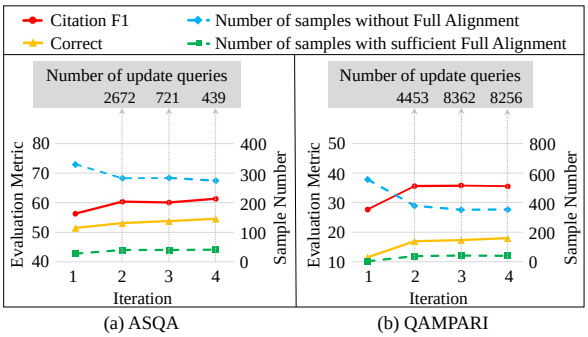
Figure 3 reveals a correlation between document category proportions and GGatrieval's performance improvements across datasets. On ELI5, performance notably improved with a higher proportion of " Partial Alignment" documents and fewer " No Alignment" ones. This is because ELI5 queries often contain redundant information, where " Partial Alignment" documents prove valuable. For instance, in the query " Please briefly explain, whether Jordan is the greatest player in NBA history," alignment with the latter segment suffices. Overall, these findings indicate that " Full Alignment" documents are most likely to support query-answer generation, " Partial Alignment" labels enhance system robustness, and " No Alignment" labels can serve as criteria for document exclusion.

Dataset	ASQA		QAMPARI	
	EM-R	Cite	Correct-F1	Cite
Origin	51.57	54.73	12.06	26.93
+SCQA	51.97	55.03	15.32	32.82
+SCQA & FGA	**52.12**	**57.16**	**16.86**	**35.57**

Table 6: Ablation Study of GGatrieval Components.

Exp-6: Ablation Study of GGatrieval Components. We further analyzed the contributions of individual GGatrieval modules on the ASQA and QAMPARI datasets. Starting from a baseline retrieval system, we incrementally added the SCQA and Fine-grained Grounded Alignment (FGA) strategies. Results in Table 6 demonstrate that each component improves performance, affirming their complementary roles in enhancing the system. Additional ablation experiments on the reflection step within the FGA strategy, detailed in Appendix B.2, further confirm its effectiveness.

Exp-7: Component Interactions Across Iterations. We investigated the interplay between query updates, document retrieval, and overall performance across four iterations using ASQA and QAMPARI datasets. Figure 4 illustrates performance metrics alongside query update number and the distribution of " Full Alignment" documents over iterations. The largest performance gain occurred after the second iteration, aligning with a decrease in samples lacking " Full Alignment" documents and an increase in those with five such documents. Performance stabilized thereafter. Notably, gains tracked closely with the rising proportion of " Full Alignment" documents, emphasizing their pivotal role. Query update frequency showed weak correlation with performance, suggesting that semantic query augmentation primarily enhances retrieval by increasing highly aligned documents. Among 1,000 samples per dataset, fewer than 30 had sufficient " Full Alignment" documents, while over 600 had none, indicating that expanding their availability in the corpus could further boost performance. These results also highlight the significant potential of interacting with retrieval documents at the level of syntactic constituent.



(a) ASQA (b) QAMPARI

Figure 4: Impact of interactions among components. The solid line represents the evaluation metric, while the dashed line indicates the distribution of Full Alignment labels within the sample.

5 Conclusion

Inspired by LLatrieval, we propose GGatrieval, a novel framework to address the semantic information deficiency issue in retrieval mechanisms. Unlike traditional methods, our approach introduces a new criterion for document selection, which are then used to classify retrieved documents. Based on this classification, we develop two strategies: FGA and SCQA. Together, these strategies optimize the retrieval mechanism, ensuring that the retrieved documents meet verifiability standards and improve overall system performance. Experimental results demonstrate that GGatrieval outperforms various types of baselines and achieves superior results.

Limitations

- We preliminarily explored interactions between queries and retrieved documents at the syntactic constituent level. However, logical relationships among these components were not modeled, and hallucination in LLMs may introduce alignment errors. Future work will focus on enhancing alignment accuracy through deep learning or reinforcement learning methods.

- Similar to human cognitive processes, our method incurs additional latency and computational overhead due to its simulation of human reasoning. However, over the years we have seen the model algorithm optimization, acceleration mechanisms advance and hardware performance increase, which can help improve the efficiency of model inference.

References

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.

Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

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

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. In *The Eleventh International Conference on Learning Representations*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023a. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023b. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

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

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Dongyub Lee, Taesun Whang, Chanhee Lee, and Heuiseok Lim. 2023. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *arXiv preprint arXiv:2309.06384*.

Leonardo Lesmo and Vincenzo Lombardo. 1992. The assignment of grammatical relations in natural language processing. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.

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

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024. LLatrieval: LLM-verified retrieval for verifiable generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5453–5471, Mexico City, Mexico. Association for Computational Linguistics.

Fengyuan Liu, Nouar AlDahoul, Gregory Eady, Yasir Zaki, Bedoor AlShebli, and Talal Rahwan. 2024. Self-reflection outcome is sensitive to prompt construction. *arXiv preprint arXiv:2406.10400*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023b. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4549–4560.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard,

Patrick Lewis, Barlas Oğuz, Edouard Grave, Wen-tau Yih, et al. 2021. The web is your oyster-knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*.

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

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. "according to . . . ": Prompting language models improves quoting from pre-training data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2301, St. Julian's, Malta. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

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

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

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2436, Toronto, Canada. Association for Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643, Mexico City, Mexico. Association for Computational Linguistics.

Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the best practices of query expansion with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1872–1883, Miami, Florida, USA. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

A Experiment Setting

A.1 Datasets and Evaluation Metrics

In this study, experiments were conducted on the ALCE benchmark (Gao et al., 2023). Given that ALCE primarily focuses on multi-hop question answering, we expanded the LLatrieval baseline by incorporating a representative open-domain single-hop QA dataset—Natural Questions (NQ) (Kwiatkowski et al., 2019)—to achieve a comprehensive evaluation of GGatrieval.

(1) ASQA (Stelmakh et al., 2022): An open-domain long-form QA dataset providing comprehensive and explanatory long-form answers to ambiguous factual questions. Answering questions from this dataset requires integrating multiple sources of information, resolving contextual ambiguities, and correlating various short answers; hence, ASQA is classified as a multi-hop dataset. We evaluate our approach using the development set of ASQA, which includes 948 questions, each with two annotations.

(2) QAMPARI (Amouyal et al., 2023): A challenging open-domain QA benchmark specifically designed to handle multi-answer questions distributed across different paragraphs. For instance, a typical question is: " Which players were drafted by the Brooklyn Nets?" Such distributed-answer scenarios frequently appear in real-world contexts. We utilize the development set of QAMPARI, comprising 1000 QA pairs, for evaluation.

(3) ELI5 (Fan et al., 2019): Developed by Facebook AI Research, ELI5 is the first large-scale open-domain dataset designed to improve AI models' capabilities in handling complex explanatory questions and generating multi-sentence, paragraph-level answers. Successfully addressing ELI5 questions requires cross-document and cross-sentence reasoning, categorizing this dataset as multi-hop QA. To assess answer correctness, we adopt the methodology from Gao et al. (2023), evaluating whether model predictions entail the sub-claims of standard answers.

(4) NQ: The Natural Questions dataset, developed by Google Research, is an open-domain QA benchmark primarily designed for evaluating machine reading comprehension, emphasizing answer localization and extraction. To maintain consistency with the ALCE evaluation, we randomly select 1000 samples from the NQ development set to assess our method.

This study assesses retrieval effectiveness through verifiability of cited documents and evaluates text generation quality of the optimized retrieval-augmented generation (RAG) system. Regarding correctness, the ASQA dataset employs Exact Match Recall (EM-R) to measure whether generated answers encompass multiple correct short answers. QAMPARI and NQ evaluates generated entity lists using precision-matched F1 scores against gold-standard answers. The ELI5 dataset assesses the entailment relationship between generated text and standard-answer claims. For verifiability, we adopt the evaluation framework proposed by Gao et al. (2023), measuring Citation Recall, Citation Precision, and their harmonic mean, Citation F1, to determine whether cited documents fully and accurately support the generated answers. This multi-dimensional evaluation approach not only provides a comprehensive assessment of model performance but also underscores the critical impact of document retrieval quality on the generated outputs. Following ALCE (Gao et al., 2023), for ASQA and QAMPARI, we use aliases of short answers provided by the dataset and normalize the model output and the short answers when measuring exact match. For ASQA, we use its sub-questions as the question to eliminate the original question's ambiguity, for simplicity.

A.2 Baselines

To achieve a comprehensive evaluation, we selected seven representative baselines from the three stages of retrieval mechanisms, highlighting GGatrieval's advantages in terms of document verifiability and overall system accuracy.

In the Pre-retrieval stage, we selected two query augmentation baselines: (1) MuGI [65], which employs LLMs to generate multiple pseudo-reference documents combined with the original query to enhance sparse and dense retrieval effectiveness; (2) Query2Doc (Wang et al., 2023), utilizing a few-shot prompting strategy with LLMs to generate pseudo-documents relevant to the query, subsequently appended to the original query to enhance its expressiveness. Specifically, we employed the BM25+MuGI (ChatGPT-3.5) method from MuGI and the Query2Doc method for query augmentation. Subsequently, the document retrieval and answer generation processes of these query augmentation baselines were kept consistent with our proposed method and evaluated on the ASQA and QAMPARI datasets. To ensure fairness, we explicitly evaluated GGatrieval's performance in the

Dataset	ASQA				QAMPARI				ELI5				Overall	
	Correct	Citation			Correct	Citation			Correct	Citation			Correct	Citation F1
	EM-R	Rec	Prec	F1	F1	Rec	Prec	F1	Claim	Rec	Prec	F1		
BM25	31.03	23.5	24.58	24.02	5.97	7.34	8.25	7.76	10.88	23.93	26.56	25.18	15.96	18.99
BGE-E-large	40.59	33.83	37.79	35.7	6.55	11.62	13.88	12.62	-	-	-	-	23.57	24.16
CRAG	33.92	33.86	36.28	35.05	5.91	7.0	8.34	7.6	11.19	29.19	32.16	30.61	17.01	24.42
RankGPT	40.17	36.59	38.68	37.61	9.24	16.8	19.7	18.13	11.2	24.89	28.52	26.58	20.20	27.44
LLatrieval	40.58	**40.13**	**42.94**	**41.5**	6.93	13.7	14.52	14.09	11.23	27.96	32.9	30.2	19.58	28.6
GGatrieval	**41.37**	39.7	42.92	41.22	**9.5**	**18.06**	**21.13**	**19.47**	**11.64**	**30.41**	**33.7**	**32.0**	**20.84**	**30.9**

Table 7: Comparison with Baselines Using the LLama Model

initial iteration of semantic compensation query augmentation, thereby eliminating potential influences from multiple iterations. **In the Retrieval stage,** we compared our approach with two standard retrievers and one trained retriever: (1) BM25 (Robertson et al., 2009), a probabilistic model widely utilized in information retrieval to evaluate relevance between queries and documents; (2) BGE Large (Liu et al., 2023b), a general embedding model efficiently converting text data into low-dimensional dense vectors, facilitating effective semantic similarity calculation and retrieval; (3) CRAG (Yan et al., 2024b), which introduces a lightweight retrieval evaluator for assessing the relevance and quality of retrieved documents given a specific query. For a fair comparison, we utilize only the retriever trained in CRAG to retrieve from the same corpus as GGatrieval. The top five ranked retrieved documents are then used to complete the final generation task. **In the Post-retrieval stage,** we selected two representative baselines: (1) RankGPT (Sun et al., 2023), directly generating document rankings through language modeling; and (2) LLatrieval (Li et al., 2024), improving retrieval quality through language model feedback, thus supporting more accurate and verifiable generation.

A.3 Implementation Details

We utilize the Verifiable Generation (Gao et al., 2023; Li et al., 2024) paradigm for both answer generation and evaluation, aiming to assess the verifiability and accuracy of the generated responses and compare the effectiveness of various retrieval mechanisms. The APIs of OpenAI's " gpt-3.5-turbo" language model and the open-source " Meta-Llama3-8B-Instruct" model are used for implementing FGA strategy, SCQA strategy, and answer generation, with the temperature set to 0 to minimize random variation. The threshold τ (Section 3.3) for document alignment categorizing is set to 0.66, determined through systematic experimentation and analysis. In the Progressive Selection (Li et al., 2024), the window size is set to 20, the number of documents retrieved per query is 5, and the number of candidate documents is 50 to ensure diversity. The number of supporting documents is set to 5, and the maximum number of iterations is 4. For ASQA, QAMPARI, and NQ datasets, the retrieval corpus is based on the Wikipedia dataset used in ALCE (Gao et al., 2023), with the dense embedding model BGE-large (Xiao et al., 2024) as the retriever. For the ELI5 dataset, we use the Sphere (Piktus et al., 2021) corpus and follow ALCE (Gao et al., 2023), employing BM25 (Robertson et al., 2009) for document retrieval due to the higher cost and slower speed of dense retrievers on large-scale web corpora. For the ALCE benchmark, the example sizes for the ASQA, QAMPARI, and ELI5 datasets are 948, 1000, and 1000, respectively. To ensure consistency with the ALCE evaluation, we randomly select 1000 samples from the development set of the NQ dataset to assess our approach.

B Supplementary Experiments and Analysis

B.1 Exp-8: Comparison with Baselines Using the LLama Model

We further compared the performance of GGatrieval with several baselines using the LLama model; results are shown in Table 7. GGatrieval generally outperformed most baseline methods, demonstrating its plug-and-play capability. Furthermore, we observed that employing stronger language models improved the performance across all methods. This suggests that GGatrieval will continue to offer practical value as increasingly powerful language models are developed in the future.

Dataset	ASQA			
	EM-R	Rec	Prec	Citation-F1
No reflection step	50.79	49.96	53.78	51.8
With reflection step	**52.86**	**56.93**	**58.19**	**57.51**

Table 8: Ablation Study of reflection step.

	ASQA	QAMPARI	ELI5
NA in all docs	10100	25736	58128
NA in final docs	1712	1915	1153
PA in all docs	6081	10715	53438
PA in final docs	1663	1594	2571
FA in all docs	2596	3386	12418
FA in final docs	1305	1337	1239

Table 9: The number of different alignment labels.

	ASQA	QAMPARI	ELI5
NA conversion rate	0.17	0.07	0.02
PA conversion rate	0.27	0.15	0.05
FA conversion rate	**0.5**	**0.38**	**0.1**

Table 10: The conversion rate of different alignment labels.

	ASQA	QAMPARI	ELI5
Total Examples	948	1000	1000
Total Docs of LLatrieval	402750	122000	126050
Total Docs of GGatrieval	18777	39837	123984

Table 11: Comparison of the number of documents retrieved by GGatrieval and LLatrieval.

B.2 Exp-9: Ablation Study on Reflection Steps

In the fine-grained semantic alignment strategy, the reflection step directly influences document alignment outcomes, subsequently determining the quality of the final candidate documents. As shown in Table 8, incorporating reflection significantly improves performance, demonstrating that the reflective capability of the LLM enhances document alignment, thus positively impacting the quality and effectiveness of the final selected documents.

B.3 Exp-10: Analysis of Alignment Label Proportions

We analyzed the quantities and conversion rates of alignment labels in the final document selection, presented in Tables 9 and 10, where " NA," " PA" , and " FA" represent " No Alignment" , " Partial Alignment" , and " Full Alignment" labels, respectively. In the ASQA, QAMPARI, and ELI5 datasets, Full Alignment and Partial Alignment documents did not dominate the final selections, primarily due to uneven label distributions and the limited availability of fully aligned documents. Nonetheless, Full Alignment documents consistently exhibited the highest conversion rates across all three datasets.

B.4 Statistics on Retrieved Documents for LLatrieval and GGatrieval

In Table 11, we present the sample sizes and the total number of retrieved documents for each dataset by LLatrieval and GGatrieval.

B.5 Further discussion

Q1. How should inference overhead be managed?

- Compared to LLatrieval, GGatrieval introduces higher latency, primarily because GGatrieval simulates human-like decision-making in document selection, whereas LLatrieval simply prompts the large model to judge document quality. Analogous to human reasoning, increased energy consumption and latency are often unavoidable, but the improved accuracy and verifiability of retrieved documents provide additional value. Notably, our approach significantly reduces the number of retrieved documents—by 95% on ASQA and by 67% on QAMPARI—which indirectly helps control inference latency.

- Our method allows dynamic control over alignment granularity via the threshold parameter and limits the maximum number of iterations (e.g., maximum iterations T=4, window size=20, and five documents retrieved per iteration), enabling a trade-off between efficiency and performance.

- As deep learning techniques and hardware continue to advance, computational overhead will become less of a concern.

Q2. How does GGatrieval contribute to multi-hop question answering?

- For multi-hop QA tasks, our method enhances the handling of cross-document and complex

semantic relations through two key mechanisms. The SCQA strategy supplements query-relevant semantic information at the level of syntactic constituents, which is particularly effective for meeting the requirements of multi-hop questions from the perspective of query subcomponents. The FGA strategy prioritizes documents that contain the greatest amount of information aligned with the query, thus supporting more accurate multi-hop reasoning.

- Another crucial factor affecting multi-hop QA performance is the generator; however, this work primarily focuses on improving the quality of retrieved documents within the retrieval mechanism.

Algorithm 1 GGatrieval

Input: Question q, document pool D_c, reranked document pool D_o, the large language model LLM, the Retriever R, the maximum iteration T, each iteration's document candidates quantity N
Output: Supporting Documents D_f

1: $Q \leftarrow q$
2: $D \leftarrow \{\}$
3: $C \leftarrow \{C_s, C_v, C_o, C_c, C_{attr}, C_{adv}, C_{supp}, C_{app}\}$
4: $\quad = LLM(I_Q, Q)$
5: **for** $i \in (1, T)$ **do**
6: $\quad$ **if** $D \neq \{\}$ **then**
7: $\quad\quad Q \leftarrow$ SCQA strategy
8: $\quad$ **end if**
9: $\quad D_c \leftarrow R(Q, N)$
10: $\quad$ **for** $D_c^* \in D_c$ **do**
11: $\quad\quad D_c^* \leftarrow$ FGA strategy
12: $\quad$ **end for**
13: $\quad D_o \leftarrow$ Rerank D_c with alignment label
14: $\quad$ **for** $D_o^* \in$ SlidingWindow(D_o) **do**
15: $\quad\quad D_f \leftarrow$ Use the LLM to select k docs from $D \cup D_o^*$
16: $\quad$ **end for**
17: $\quad$ **if** Verify(q, D_f) $\rightarrow$ Yes **then**
18: $\quad\quad$ **break**
19: $\quad$ **end if**
20: **end for**
21: **Return** D_f

C Algorithm of GGatrieval

Algorithm 1 outlines the workflow of GGatrieval. The process begins with parsing the user query into syntactic constitutents $C = \{C_s, C_v, C_o, C_c, C_{attr}, C_{adv}, C_{supp}, C_{app}\}$ (Line 3∼4), initiating the iterative process. In each iteration, the system applies the SCQA strategy to retrieve a refined set of candidate documents D_c (Lines 6∼9), ensuring that the retrieved documents are increasingly semantically aligned with the query. Next, the FGA strategy is employed to assign alignment labels to each document in D_c (Lines 10∼11). These documents are then reordered based on the alignment labels and relevance, resulting in a prioritized set D_o (Line 13), which includes documents that meet the verifiability criteria. Finally, the system employs the Progressive Selection and Document Verification methods proposed by Li et al. (2024) to select and validate the final supporting documents D_f(Lines 14∼18). GGatrieval defines a robust selection criterion to establish clear retrieval objectives. Through the iteration, the retrieval results are progressively refined to yield documents that better align with the retrieval goal, thereby enabling the LLM to generate both accurate and verifiable answers.

D Instructions of GGatrieval

We show the overall instructions in Table 12, 13, 14, 15. The instructions for the progressive selection process are identical to those used in LLatrieval(Li et al., 2024).

CONTEXT
Profile
You are a linguistics expert proficient in English grammar. You want to analyze the grammatical components of a question.

Skill
Analyze the grammatical structure of the given question from the perspectives of the subject, predicate, object, attribute, adverbial, complement, etc.

OBJECTIVE
From the perspective of grammatical structure such as subject, predicate, object, attribute, adverbial or complement, please parse the given question grammatically and return it in a standard format.
<question>
Question
</question>

Output
Just output the syntactic components of the given question according to the standard format, do not output any other content.

Output Criteria (Very Important)
Be as objective as possible.

STYLE
Please generate specific content in a very rigorous style, following the writing habits of a linguistics professor.

REPONSE
Standard format of syntactic components to the question.

Table 12: The instruction for syntactic parsing.

CONTEXT
Profile
You are a linguistics expert proficient in English grammar.You want to find the answer to a question in a piece of text, but you are not sure if the text contains the answer to the question.

Skill
1. Analyze the grammatical structure of the given text from the perspectives of the subject, predicate, object, attribute, adverbial, complement, etc.
2. According to all grammatical components of the question to find the corresponding content that matches or indicates in semantics in the given text.

OBJECTIVE
From the perspective of grammatical structure such as subject,predicate object, attribute, adverbial or complement,etc, please judge whether the given text has sufficient content to semantically match or indicate each syntactic component of the given question.And give your analysis steps.

<question>
Question
</question>

<syntactic component>
Components
</syntactic component>

<text>
Passage
</text>

Rules
1. Output specific analysis steps.
2. Assume you do not know the answer to the question, and analyze and judge based solely on the content of the given text.
3. Strictly follow the specified output format. Do not answer the given question.

Output
-analysis steps.
-Judgement Result.

Workflow
1. Analyze the text to find content semantically matches or indicates for each syntactic component of the question.
2. Make an analysis result for each syntactic component.

STYLE
Please generate specific content in a very rigorous style, following the writing habits of a linguistics professor.

REPONSE
Analysis steps and results.

Table 13: The instruction for Fine-grained Grounded Alignment.

CONTEXT
Now there is a question, a syntactic components list for that question, a given text, and an analysis result of semantic matching between the text and the question.I want to reflect on the given analysis results and output a new list. Each element in the new list comes from the syntactic components list and can be semantically matched or indicated with content from the given text.

OBJECTIVE
Please reflect on whether the analysis results are correct, provide the correct analysis with a conclusion again.For each element in the syntactic components list, if it can find semantically matching or indicating content in the given text, please put that element into a new list and output this new list,else,and rewrite a more specific question by converting the missing components into synonymous descriptions.

<question>
Question
</question>

<syntactic component>
Components
</syntactic component>

<analysis results>
Analysis_Results
</analysis results>

<text>
Passage
</text>

Output
-Analysis Steps:Correct analysis with a conclusion.
-Judgement Result:The syntactic components list.
-Rewrite Question:The question is rewritten by converting the missing components into synonymous descriptions,and enclose it in " «<" and " »>" symbols.

Rules
1. Output specific correct analysis steps with a conclusion.
2. Each element in the final output list in the Judgement Result must be able to find semantic match or indication in the given text.
3. Assume you do not know the answer to the question, and analyze and judge based solely on the content of the given text.
4. Strictly follow the specified output format. Do not answer the given question.
5. The final output list must start with '[' and end with ']'.
6. Each element of the final output list in Judgement Result must come from the syntactic components list,otherwise output a empty list.
7. Enclose the Rewrite Question in " «<" and " »>" symbols.

REPONSE
1. Analysis steps with a conclusion.
2. The final output list as Judgement Result,without any other content.
3. A Rewrite Question enclosed with " «<" and " »>" symbols.

Table 14: The instruction for Reflection and Query Optimization.

OBJECTIVE
Please generate a paragraph related to the given question, such that for every grammatical component of the question, there is a semantically matching grammatical component in the paragraph, and the paragraph can provide an answer to the question.

DEMONSTRATION
Who directed a movie written by Ken Hixon?//relative paragraph:Fear and Loathing in Las Vegas (film) Fear and Loathing in Las Vegas is a 1998 American psychedelic satirical road film adapted from Hunter S. Thompson's novel of the same name. It was co-written and directed by Terry Gilliam, starring Johnny Depp as Raoul Duke and Benicio del Toro as Dr. Gonzo. The two embark on an initially assigned journey with journalistic purpose which turns out to be an exploration of the Las Vegas setting under the effect of psychoactive substances. The film received mixed reviews from critics and was a financial failure.

Who's job is in the LA County Sheriff's Department?//relative paragraph:Jim McDonnell (sheriff) James McDonnell (born 1959) is an American law enforcement official who served as the 32nd Sheriff of the County of Los Angeles in California. McDonnell was elected as L.A. County's 32nd sheriff on November 4, 2014, defeating former Undersheriff Paul Tanaka. He replaced interim sheriff John Scott on December 1, 2014, when he was sworn in. Previously he served as the Chief of Police in Long Beach, California and before that in the Los Angeles Police Department, reaching the rank of Assistant Chief. McDonnell grew up in a working-class neighborhood in Brookline, Massachusetts.

Who worked for a military branch of the Kingdom of Prussia?//By the end of Frederick's reign, the army had become an integral part of Prussian society and numbered 200,000 soldiers, making it the third largest in Europe after the armies of Russia and Austria. The social classes were all expected to serve the state and its army — the nobility led the army, the middle class supplied the army, and the peasants composed the army. Minister Friedrich von Schrötter remarked that, " Prussia was not a country with an army, but an army with a country" . Frederick the Great's successor, his nephew Frederick William II (1786–97).

What Indonesian mosques are located in the province of South Sulawesi?//However, such concerns were allayed along with the development and progress of the renovations since the groundbreaking by then governor of Palembang Zainal Basri Palaguna in October 9, 1999. Great Mosque of Makassar Great Mosque of Makassar is a mosque located in Makassar, Indonesia, and the main mosque of South Sulawesi Province. The construction begun in 1948 and completed in 1949. Since then the mosque underwent a renovation from 1999 to 2005. The mosque can accommodate up to 10,000 worshipers, making it one of the largest mosques in Southeast Asia.

{Question}

Table 15: The instruction of generating semantically aligned pseudo-documents.