Momoka-RAG: MCTS-Organized Mapping of Knowledge Associations for Long-Document Retrieval Augmented Generation

Anonymous ACL submission

Abstract

Existing frameworks remain trapped in a passive and mechanical approach in constructing knowledge structure, which only allows them to uncover superficial associations between chunks while lacking proactive exploration of deeper semantic relationships among them. To address the aforementioned issues, we propose Momoka-RAG (MCTS-Organized Mapping of Knowledge Associations for Long-Document Retrieval Augmented Generation). It employs the **Momoka-Map** to utilize Monte Carlo Tree Search (MCTS) to proactively uncover connections among chunks and construct optimal semantic information paths with the objective of completing semantic relationships. On this basis, the **Momoka-Trail Retriever** further expands and filters the chunk candidate pool to retrieve the chunks most relevant to the query. Experiments on datasets including Dragonball, SQUAD, NFCORPUS, SCI-DOCS, HotpotQA, and TriviaQA demonstrate that for long-document retrieval tasks, our framework achieves higher precision while maintaining competitive recall compared to other RAG frameworks.

1 Introduction

The RAG framework represents a groundbreaking paradigm in the field of natural language processing (Lewis et al., 2021), designed to address several inherent limitations of LLMs - such as hallucination, outdated knowledge, and inadequate domain-specific adaptation (Augenstein et al., 2023) - by incorporating external knowledge.

In recent years, LLMs have evolved beyond simple QA to more complex tasks involving long documents, making RAG essential for accurately retrieving relevant passages.

Nowadays, mainstream RAG techniques no longer rely solely on fixed-length chunking methods. For instance, Meta-Chunking (Zhao et al., 2024) reformulates the chunking task as a binary

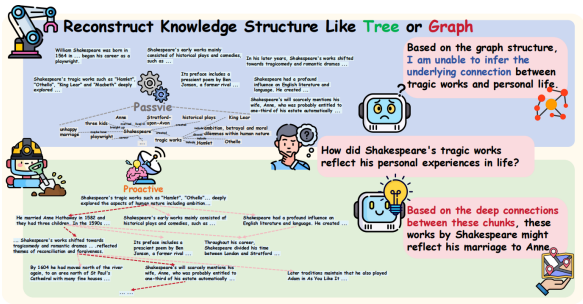

Figure 1: In long documents, existing frameworks like GraphRAG and RAPTOR fall into **passive and mechanical** constraints when constructing knowledge structures, limiting their ability to uncover anything beyond **superficial and localized relationships between chunk**s. Therefore, a semantic-completion-oriented **active exploration** mechanism is needed to **excavate deeper semantic connections between chunks.**

classification problem for LLMs. Dense X Retrieval (Chen et al., 2024) decomposes original document into minimal semantic units called "propositions". In contrast, Late-Chunking (Günther et al., 2024) innovatively inverts the chunking process by adopting an "embed first, then chunk" strategy.

Other RAG frameworks have approached the problem by reorganizing knowledge structures to establish connections between chunks. For instance, RAPTOR (Sarthi et al., 2024) employs Gaussian mixture clustering combined with LLM-generated summaries for each cluster, thus constructing a hierarchical tree structure. Knowledge graph-based frameworks utilize LLMs to extract entities and relationships from documents, thereby building graph-based representations.

However, the fundamental reason why RAG frameworks that rely solely on chunking methods are unsuitable for long-document scenarios lies in their failure to strike a balance between maintaining local coherence of chunks and preserving cross-paragraph semantic connections.

Although RAPTOR and graph-based RAG have

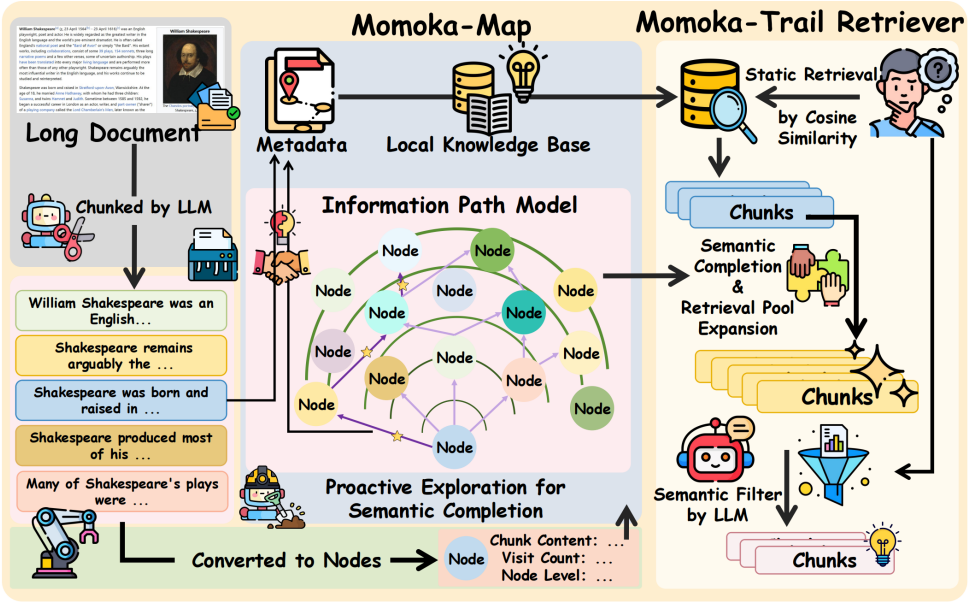

Figure 2: Framework of Momoka-RAG.

made initial progress in establishing inter-chunk relationships, their approaches to knowledge structure reconstruction remain mechanistic and passive, often only capturing local and superficial associations between chunks while struggling to uncover deeper semantic connections. Specifically, RAPTOR relies on unsupervised clustering to construct hierarchical structures — a method that is inherently passive and statistically driven. Its associations are based on vector similarities between chunks rather than genuine semantic reasoning, treating all chunks as equally significant during the clustering process and generating undifferentiated summaries. This results in a rigid tree structure that fails to actively identify and validate the profound connections and logical threads between superficially similar chunks. On the other hand, graph-based RAG performs undifferentiated entity and relationship extraction on all chunks, where the quality of associations heavily depends on hyperparameters such as window size. The extraction method also limits its ability to capture only local and superficial relationships, ultimately constraining the graph structure's capacity to model complex semantic relationships and long-range dependencies. Consequently, the constructed association network lacks semantic richness and depth in reasoning.

Therefore, to address the aforementioned issues and explore how to proactively mine deep-seated semantic relationships between chunks guided by semantic completion, we propose Momoka-RAG, which consists of two components: Momoka-Map and Momoka-Trail Retriever.

During the construction of the Momoka-Map component, we draw inspiration from the MCTS method (Browne et al., 2012). Traditionally, when MCTS is applied to RAG systems, it is commonly used to optimize the retrieval process — as seen in frameworks like MCTS-RAG (Hu et al., 2025), where each retrieval step is treated as a node. However, Momoka-Map adopts a different perspective: we treat each chunk segmented by the LLM as a node and, guided by the goal of complementing semantic information, use MCTS to explore the connections between chunks, thereby constructing a semantically enriched information path model.

Building upon the information path model provided by the Momoka-Map component, we further designed the Momoka-Trail Retriever. This component first efficiently retrieves relevant chunks through static similarity retrieval, then leverages the path model to complement semantic information. Simultaneously, it incorporates the powerful semantic reasoning capabilities of LLM to further expand and filter the candidate pool based on the user's query.

In Momoka-RAG, we shift the application of MCTS from the retrieval phase to the indexing phase. This design effectively mitigates the poor real-time performance typically associated with MCTS methods, avoiding the high latency issues that would arise if applied during retrieval. At the same time, it allows the retrieval component to ben-

efit from the precomputed path model, significantly improving retrieval efficiency.

To demonstrate the effectiveness of our proposed framework, we conducted experiments on carefully selected datasets including Dragonball (Zhu et al., 2025), SQUAD (Rajpurkar et al., 2016), NFCORPUS (Boteva et al., 2016), SCI-DOCS (Cohan et al., 2020), HotpotQA (Yang et al., 2018), and TriviaQA (Joshi et al., 2017). Evaluation metrics such as Recall@k (Musgrave et al., 2020), MRR@k (Voorhees and Tice, 2000), Precision@k, EIR@k, F1@k, METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), time latency, as well as human evaluation metrics specifically proposed for the Dragonball dataset, were employed. Experimental results indicate that our framework achieves competitive performance in long-document retrieval tasks compared to existing approaches.

The main contributions of this work are as follows:

(1)We propose Momoka-Map, which applies MCTS to construct a semantically enriched information path model while shifting the high computational cost of MCTS to the pre-processing indexing phase.

(2) We introduce the Momoka-Trail Retriever, which leverages the information path model and incorporates the powerful reasoning capabilities of LLM to efficiently retrieve and filter chunks.

(3) We demonstrate competitive performance on carefully selected datasets including Dragonball, SQUAD, NFCORPUS, SCI-DOCS, HotpotQA, and TriviaQA.

2 Related Work

Today's LLMs have demonstrated remarkable reasoning capabilities and the ability to process extremely long contexts. However, when applied to professional domain-specific question answering, they often produce factual inaccuracies due to a lack of relevant training data (Zhao et al., 2025). As a result, it becomes necessary to provide relevant knowledge alongside the user's query. Yet, as the application scope of LLMs continues to expand, tasks involving long-document processing are increasingly common. In such scenarios, directly inputting the entire document content along with the user's question is no longer feasible, since the length of these documents often exceeds the practical context window of the models and introduces significant noise. Therefore, RAG technology has become particularly critical.

Naive RAG represents the most fundamental and widely used paradigm in RAG. This framework typically segments documents into fixed-length chunks and employs methods such as DPR, BM25, or vector similarity search to retrieve the top-k most similar chunks using k-nearest neighbors (KNN) (Cover and Hart, 2006) algorithm. For instance, the LangChain library (Chase, 2024) offers various text splitters and embedding methods to facilitate this process. While this paradigm is well-suited for shorter documents with concentrated information points, it often leads to issues such as contextual fragmentation, redundancy, and introduced noise when handling long documents.

To mitigate the issue of chunk fragmentation, subsequent research has proposed more advanced segmentation strategies. Meta-Chunking leverages LLMs to generate chunks of flexible length and coherent content by evaluating per-sentence perplexity or calculating probability differentials for segmentation decisions, making it more suitable for processing well-structured documents. Dense X Retrieval, on the other hand, decomposes text into finer-grained propositional units that encapsulate individual fact fragments. However, these methods remain inadequate for long documents because, despite improving chunk quality, they still fail to establish semantic relationships between chunks.

Late-Chunking offers a straightforward yet effective approach to mitigate the isolation of chunks by ensuring each chunk retains contextual information from its surroundings. It adopts an "embed-first, then chunk" strategy, which preserves both the granularity of chunks and the integration of contextual information into their embedding vectors. This method demonstrates particular strength in processing medium-length documents. However, when applied to long documents, the increasing number of tokens leads to information dilution at the embedding level, ultimately compromising retrieval performance.

In order to establish connections between chunks, mainstream RAG frameworks have begun to reorganize knowledge structures. RAPTOR treats chunks as leaf nodes and employs Gaussian mixture clustering combined with the summarization capability of LLMs to construct a tree structure in a bottom-up manner. However, when applied to long documents, the large number of chunks leads to increased hierarchical levels, inevitably caus-

ing loss of detail in higher-level summaries. On the other hand, during the clustering process, all chunks are treated as equally important and summarized indiscriminately, which amplifies the shortcomings of the inherently rigid tree-like structure in long-document scenarios.

Knowledge graph-based frameworks such as GraphRAG (Edge et al., 2024), LightRAG (Guo et al., 2025) and HippoRAG (Gutiérrez et al., 2025) construct graph structures by extracting entities and relationships from chunks. However, they primarily capture local and superficial associations within chunks, lacking a purpose-driven reasoning process aimed at semantic completion. This limitation prevents them from actively uncovering deeper semantic connections between chunks across documents—a critical shortcoming especially prominent in long-document scenarios.

3 Momoka-RAG

In this section, we will introduce how Momoka-Map incorporates MCTS into the RAG framework to derive a semantic information path model, as well as how Momoka-Trail Retriever leverages this model to achieve efficient chunk retrieval and filtering.

3.1 Momoka-Map

Mainstream RAG frameworks often fall into two contrasting yet equally deficient paradigms when addressing semantic associations in long document. On one hand, approaches represented by RAPTOR and GraphRAG focus on constructing knowledge structures in a mechanistic and passive manner, which limits their ability to capture only local and superficial relationships between chunks. On the other hand, another line of research concentrates on dynamically optimizing the retrieval workflow itself through active exploration, yet without fundamentally reorganizing the underlying knowledge structure. This led us to explore whether it is possible to integrate active exploration into the construction of knowledge structures. Inspired by MCTS, we treat each chunk as a node and, guided by the goal of semantic completion, proactively explore the optimal information path for each chunk to build a corresponding model. This shifts the focus from "which chunks are closer in the vector space" to "which chunks should be retrieved next to fully understand the meaning of this chunk?".

The Momoka-RAG framework begins with an LLM-driven document parsing and segmentation phase. Given a long document D, we leverage the deep linguistic understanding capabilities of a LLM to dynamically decompose it into a sequence of semantically coherent chunks. This process can be formally defined as follows:

$$C = F_{Segment}(D) = \{c_1, c_2, ..., c_N\}, \quad (1)$$

where each text chunk c_i is a triple:

$$c_i = (text_i, p_i, s_i), \quad (2)$$

Here, $text_i$ represents the textual content of the i-th chunk, while p_i and s_i denote the paragraph number and the sentence number of this chunk in the original document D respectively.

Building upon this foundation, the core task of Momoka-Map is to construct a semantic information path model P_i of each chunk c_i in the sequence C. To achieve this goal, we designed the following MCTS framework guided by semantic completion, treating each chunk c_i as a node:

(1) **Selection**: Starting from the root node i, recursively select the child nodes until a leaf node is reached. To strategically choose the branch most likely to yield high rewards, we balane the core trade-off between "exploitation" and "exploration". The selection strategy follows the Upper Confidence Bound for Trees (UCT) (Kocsis and Szepesvari, 2006) formula:

$$UCT(i) = \frac{R_i}{N_i} + C \cdot \sqrt{\frac{\ln N_p}{N_i}}, \quad (3)$$

where R_i is the cumulative reward of node i, N_i is its visit count, N_p is the visit count of its parent node, and C is the exploration constant.

(2) **Expansion**: When a leaf node is selected and is not fully explored - meaning there exist nodes that have not yet appeared in the current path - the algorithm randoml selects one such unexplored node c_{new} as a child node for expansion. This can increase the breadth of exploration while avoiding path loops.

(3) **Simulation**: Starting from the currently expanded node, a random rollout is performed. In traditional MCTS simulations, the rollout process only considers the outcome of the final node to determine the reward. However, in the task of "semantic completion," following this rollout logic would imply that only the last node contributes meaningfully, which clearly contradicts the objective of the

task. Therefore, we design the rollout process such that each step contributes to the reward, ensuring that every expanded node along the path participates in evaluating the semantic comlementation effect.

At each step of the rollout, an LLM acts a a "semantic referee" to make a binary judgment on whether c_{new} contributes to the semantic completion of the root node c_i. Simultaneously, incorporating structural priors, we design a dedicated evalution function:

$$\begin{aligned} r = \mathbb{I}_{LLM}(c_i, c_{\text{new}}) \\ + \alpha \cdot \frac{1}{|p_i - p_{\text{new}}| + \gamma} \\ + \beta \cdot \frac{1}{|s_i - s_{\text{new}}|} + \delta, \end{aligned} \tag{4}$$

where $\mathbb{I}_{LLM}(\cdot)$ is an indicator function that outputs 1 if determines a semantic completion relationship exists between the two chunks, and o otherwise. α and β are weighting coefficients, while γ and δ are smoothing constants. This function design ensures that the evaluation reward r captures both deep semantic associations and the structural logic of the document. The specific prompt for using an LLM to determine semantic completion relationships can be found in Appendix A.

In this specific task, there is no termination condition such as "win/loss" as in traditional MCTS applications like board games. Nor can we allow the rollout to continue until all nodes are simulated, as doing so would be equivalent to retrieving the entire document, defeating the purpose of efficient retrieval. Therefore, we set a maximum rollout step limit $N_{rollout}$ as the termination condition. The total reward R for the entire rollout path is computed as the average of all step rewards, serving as the final return for the current simulation:

$$R = \frac{1}{L_{rollout}} \sum_{k=1}^{L_{rollout}} r_k, \tag{5}$$

(4) **Backpropagation**: The reward R obtained from the simulation is backpropagated along the search path, updating the cumulative reward R_i and visit count N_i for all nodes on the path.

After a substantial number of iterations, the algorithm starts from the root chunk c_i and selects the child node with the highest visit count as each level, proceeding recursively until an optimal information path model $P_i = (c_i, c_j, c_k, ...)$ is formed. This sequence of chunks is considered the most semantically coherent "storyline" that best complements the contextual meaning of the root chunk c_i. Ultimately, for all chunks in the document, Momoka-Map outputs a set $\mathcal{G} = P_1, P_2, ..., P_N$, which is associated with the chunk sequence C and then stored together in the vector database.

Algorithm 1: Momoka-Trail Retriever

Input: Query Q, Retriever $\mathcal{R}$, Reranker $\mathcal{S}$, LLM $\mathcal{M}$, Semantic Information Path $\mathcal{P}$, Top-k K
Output: Retrieved set $\mathcal{F}$
$C_{init} \leftarrow \mathcal{R}$.get_relevant_chunks$(Q, K)$
$C_{aug} \leftarrow \emptyset$
for $c \in C_{init}$ **do**
  $P_c \leftarrow \mathcal{P}(c)$ `// Get Corresponding information path`
  for $p \in P_c$ **do**
    $C_{aug} \leftarrow C_{aug} \cup \{c \oplus p\}$
  end for
end for
$\mathcal{E}_{LLM_{relevant}} \leftarrow \emptyset$ $\mathcal{E}_{LLM_{irrelevant}} \leftarrow \emptyset$
for $c_{aug} \in C_{aug}$ **do**
  $p \leftarrow$ "`Determine relevance: // Knowledge:` c_{aug} `// Question:` Q `// Output: 1/0`"
  $r \leftarrow \mathcal{M}(p)$
  if $r = 1$ **then**
    $\mathcal{E}_{LLM_{relevant}} \leftarrow \mathcal{E}_{LLM_{relevant}} \cup \{c_{aug}\}$
  else
    $\mathcal{E}_{LLM_{irrelevant}} \leftarrow \mathcal{E}_{LLM_{irrelevant}} \cup \{c_{aug}\}$
  end if
end for
if $|\mathcal{E}_{LLM_{relevant}}| \geq K$ **then**
  $\mathcal{F} \leftarrow \mathcal{S}(Q, E_{LLM_{relevant}}, K)$
else
  $\mathcal{F} \leftarrow \mathcal{E}_1 \cup \mathcal{S}(Q, E_{LLM_{irrelevant}}, K - |\mathcal{E}_1|)$
end if
return $\mathcal{F}$

Description: $\mathcal{R}$.get_relevant_chunks(Q, K) refers to the operation of retrieving the top-K chunks with the highest similarity scores from the vector database in response to the query Q. $\mathcal{S}(Q, E, K)$ refers to the operation of re-ranking a set of E chunks according to the query Q and then selecting the top-K chunks from the re-ranked list.

3.2 Momoka-Trail Retriever

Momoka-Map constructs high-quality semantic information paths for each chunk in the document. The role of Momoka-Trail Retriever is to dynamically leverage these precomputed paths during retrieval, screening the most relevant context for a query from a broader and more semantically enriched candidate pool. Its core idea is to treat not individual chunks as independent retrieval units, but rather the entire semantic path to which a chunk belongs as an expandable candidate resource. This

retriever employs path expansion and intelligent refinement to efficiently retrieve relevant chunks.

Given a user's query q, it is first embedded into a vector. Using cosine similarity, the $Top - k$ most relevant initial text chunks are retrieved from the vector database created by the Momoka-Map component, denoted as $V_q = \{v_1, v_2, ..., v_k\}$.

For each initially retrieved text chunk v_i, its pre-computed semantic information path P_i is loaded. Each related chunk $c_j \in P_i$ is then concatenated with v_i to form a series of semantically enriched augmented chunks $\{e_{i_1}, e_{i_2}, ...\}$. All such augmented chunks constitute the candidate pool E.

Each enriched chunk from E, along with the query q, is fed into an LLM. A binary classification task is performed to assess its relevance at the contextual relevance level. Only chunks classified as relevant are retained, forming the refined set V_{filtered}. Finally, the refined set V_{filtered} is re-ranked using a reranker model based on semantic similarity to produce the final retrieved set V_{final}. In Algorithm 1, we show this retrieval strategy.

4 Experiments

Datasets. Experiments are conducted on four datasets: Dragonball, SQUAD, NFCORPUS, SCIDOCS, HotpotQA and TriviaQA, filtered by document length. Only the Finance subset of Dragonball is used, as other subsets contain structured content like legal judgments and medical records, not coherent text.

Embedding Model and Reranker. Since both the proposed framework and the baseline methods used for comparison are not dependent on specific embedding models, and switching embedding models does not significantly affect the functionality or ranking performance of the frameworks, we consistently employ BGE-M3 (Chen et al., 2023) - which demonstrates strong performance across multiple languages and domains - as the embedding model in all experiments. The $batch_size$ was set to 32, and $normalize_embeddings$ was set to True, meaning generated embedding vectors were normalized. In the experiments, we use the bge-reranker-large as the reranker model. The parameters of the aforementioned model all adopt the default parameter settings of the BCERerank function in the BCEmbedding repository [1].

LLM. For tasks involving the use of pre-trained LLMs for Chunkink, filtering, summarization, information extraction, and answer generation, we employ the OpenAI API to access the highly capable Qwen-max (Bai et al., 2023) model with default parameter settings. In the generative quality evaluation phase, we switch to DeepSeek-R1 (DeepSeek-AI et al., 2025) to perform the task, still using the OpenAI API with default parameters. For Meta-Chunking and Dense X Retrieval, we utilize the models recommended in their respective code repositories — specifically, Qwen2.5-1.5B-Instruct and propositionizer-wiki-flan-t5-large — and load these models using the AutoTokenizer and AutoModelForCausalLM interfaces from the transformers library under default parameter configurations.

Chunks Size. To ensure experimental consistency, for frameworks that require manually inputting chunks — such as Momoka-RAG, LightRAG, and RAPTOR — we use chunks segmented by the LLM. For frameworks involving expected chunk size settings, such as Meta-Chunking, we configure the size according to the average length of the LLM-segmented chunks to maintain uniformity. All other frameworks are maintained their default configurations. The average chunk length for each dataset is shown in Appendix C.

Metrics. For retrieval evaluation metrics, we adopt Recall, Precision, EIR, MRR, and F1 as the main evaluation indicators. For generation evaluation metrics, we use ROUGE-L, METEOR, BLEU, Wrong, Irrelevant, and Relevant as the evaluation criteria. The final metric score is calculated using Formula 6.

$$Metric = Metric@1 + Metric@3 + Metric@5 \tag{6}$$

Hyper-parameter. Through experimental verification and comprehensive consideration of factors such as performance and resource consumption, we set the hyper-parameters as follows in our experiments. The MCTS iteration count was fixed at 100, and the maximum rollout length was limited to 5. Additionally, α is set to 3, β is set to 2, and γ and δ are set to 1.

Statistical Significance. For retrieval metrics, we employed permutation tests for verification; for generation quality metrics, we adopted stratified sign tests for validation. We use ★ ($p < 0.05$) to mark statistically significant wins.

[1] https://github.com/netease-youdao/BCEmbedding

Methods	ROUGE-L↑	METEOR↑	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	Relevant↑	Irrelevant↓	Wrong↓
Momoka-RAG	0.286	0.333	0.190	0.136	0.104	0.083	64.870	32.949	2.182
LightRAG	0.232	0.294	0.168	0.113	0.083	0.064	48.872	46.740	4.388
PPL	0.268	0.305	0.144	0.107	0.083	0.066	74.097	23.621	2.282
MSP	0.282	0.323	0.181	0.131	0.101	0.080	65.095	32.121	2.783

Table 1: **Comparative Experiments on Generation:** Only the Dragonball dataset provides human-annotated detailed answers, so we only conduct generation quality experiments on it. In the table, **PPL** stands for Meta-Chunking-PPL.

Methods	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
LightRAG	5.775e-46 ★	5.425e-29 ★	7.089e-22 ★	3.13e-30★	6.433e-32★	2.916e-32 ★
PPL	4.0308E-25★	1.337E-25★	4.205E-43★	9.869E-27★	1.431E-19★	7.399E-16★
MSP	0.235001	0.00075★	0.00040★	1.21E-05★	0.00011★	0.0084★

Table 2: **Statistical Validation Experiments on Generation Quality:** This table presents the results of the statistical significance comparison between SAKI-RAG and the other methods on the text generation quality metrics. The values in the table are p-values, where the symbol ★ indicates a statistically significant difference compared to SAKI-RAG (e.g., $p < 0.05$). In the table, **PPL** and **MSP** stand for Meta-Chunking-PPL and Meta-Chunking-MSP.

4.1 Comparative Experiments on Retrieval Performance

In terms of retrieval performance, we compare our proposed Momoka-RAG against popular RAG frameworks—including Late-Chunking, RAPTOR, Meta-Chunking-PPL, Meta-Chunking-MSP, and Dense X Retrieval. The retrieval performance results are presented in Table 3. Due to space limitations, only the F1, Precision, and MRR metrics are included in the main text; complete experimental results can be found in Appendix D.

In the experiments of this subsection, we compare Momoka-RAG with retrieval-oriented RAG frameworks such as Late-Chunking, RAPTOR, Meta-Chunking-PPL, Meta-Chunking-MSP, and Dense X Retrieval. In the results table, the top two best-performing frameworks for each metric are highlighted in blue, with a darker shade indicating the first-ranked method and a lighter shade representing the second-ranked method. From the experimental results, Momoka-RAG demonstrates outstanding performance in F1, Precision, and MRR metrics, achieving the highest scores in most cases. This indicates that, compared to other frameworks, Momoka-RAG effectively balances accuracy and completeness while efficiently prioritizing the most relevant chunks. We also conduct retrieval metric experiments on LightRAG, as shown in the complete table in the Appendix D.

The result tables of other metrics such as Recall, EIR and Time Latency can be found in Appendix D. Regarding the Recall metric, although Momoka-RAG does not always achieve the top score, it still delivers a decent performance. On the EIR metric, Momoka-RAG does not score well, indicating a certain degree of redundancy in the retrieved chunks. However, our subsequent experiments on generation quality demonstrate that this level of redundancy does not compromise the accuracy of the generated answers. However, the introduction of an LLM for filtering during the retrieval stage incurs a computational cost. The total time consumption increases with the Top-K value, but it remains within acceptable limits.

To verify that the observed performance differences are substantial and reliable, we conduct tests of statistical significance. The results are presented in Table 4 (Due to space limitations, only the Top-5 results are reported in the main text; the complete data can be found in Appendix F.). As shown, SAKI-RAG demonstrate statistically significant differences on the vast majority of metrics. This confirms that its superior performance is not coincidental but is statistically reliable.

The aforementioned experimental results and analysis demonstrate that Momoka-RAG's strategy of "proactively exploring inter-chunk connectivity guided by semantic completion" achieves strong performance in long-document retrieval tasks.

4.2 Comparative Experiments on Generation Quality

For generation quality assessment, we used automated metrics alongside the LLM evaluation prompts from the Dragonball dataset (metrics: **Relevant**, **Irrelevant**, **Wrong**). We compared Momoka-RAG against LightRAG and Meta-Chunking, with results in Table 1. As evidenced

Methods	Dragonball			SQUAD			NFCORPUS			SCI-DOCS			HotpotQA			TriviaQA		
	F1↑	Pre.↑	MRR↑	F1↑	Pre.↑	MRR↑	F1↑	Pre.↑	MRR↑	F1↑	Pre.↑	MRR↑	F1↑	Pre.↑	MRR↑	F1↑	Pre.↑	MRR↑
Momoka-RAG	1.18	218.21	2.28	2.84	279.34	2.88	2.78	277.48	2.79	2.80	279.25	2.79	2.84	283.24	2.85	0.81	81.47	0.81
Late-Chunking	0.03	3.51	0.04	0.43	34.58	0.45	0.07	4.79	0.11	0.02	1.02	0.03	0.10	7.81	0.12	0.47	46.72	0.47
RAPTOR	0.71	101.08	1.37	/	/	/	2.73	261.30	2.70	2.83	275.10	2.85	/	/	/	/	/	/
PPL	1.16	133.86	2.20	1.61	133.26	2.47	2.62	243.53	2.79	2.74	264.17	2.83	0.76	69.07	0.82	0.14	14.29	0.01
MSP	0.88	111.39	1.85	1.64	134.77	2.49	2.66	251.84	2.75	2.78	266.64	2.86	2.56	237.60	2.80	0.58	57.92	0.15
DXR	0.03	6.14	0.11	1.19	96.24	1.68	2.73	264.79	2.74	2.79	270.14	2.83	1.42	118.78	2.30	0.38	37.83	0.38

Table 3: **Comparative Experiments on Retrieval:** Due to the presence of sensitive or unsafe content in the original documents of datasets, LLMs cannot be used to build tree structures for RAPTOR. Furthermore, as the value of Top-k increases, chunks containing sensitive information are retrieved, causing the LLM component responsible for the filtering function in Momoka-RAG to fail. Therefore, in the TriviaQA experimental results, we only present the Top-1 metric results. In the table, we abbreviate the metrics and framework names where **Pre.**, **PPL**, **MSP** and **DXR** stand for Precision, Meta-Chunking-PPL, Meta-Chunking-MSP and Dense X Retrieval.

Methods	Dragonball		SQUAD		NFCORPUS		SCI-DOCS		HotpotQA		TriviaQA	
	Pre.	F1	Pre.	F1	Pre.	F1	Pre.	F1	Pre.	F1	Pre.	F1
RAPTOR	3.3E-05★	3.3E-05★	/	/	3.3E-05★	3.3E-05★	2.0E-04★	1.6E-01	/	/	/	/
PPL	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★
MSP	3.3E-05★	3.3E-05★	1.3E-03★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	5.3E-04★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★
DXR	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★	3.4E-03★	2.3E-04★	9.0E-02	3.3E-05★	3.3E-05★	3.3E-05★	3.3E-05★

Table 4: **Statistical Validation Experiments on Retrieval:** This table presents the results of the statistical significance comparison between SAKI-RAG and the other methods on retrieval metrics. In the table, **Pre.**, **PPL**, **MSP** and **DXR** stand for Precision, Meta-Chunking-PPL, Meta-Chunking-MSP and Dense X Retrieval.

by the results, Momoka-RAG achieves superior generation performance compared to LightRAG while maintaining a simpler framework. In contrast to MSP and PPL, despite setting a target size for them to ensure experimental fairness, the final generated chunks still exceed this predefined value. This allows them to cover more contextual content but also leads to relatively higher Wrong metric. On the other hand, Momoka-RAG functions as a plug-and-play framework that can be applied in subsequent stages of chunking methods like PPL and MSP, further enhancing their performance. For details regarding the corresponding prompts and metric descriptions, please refer to Appendix E.

To determine whether the observed performance differences are substantial or not, we conduct statistical significance tests, with the results presented in Table 2. As shown, SAKI-RAG demonstrate statistically significant differences across vast majority of metrics. This provides strong evidence that performance improvement brought by SAKI-RAG framework is statistically reliable.

5 Conclusions

In this paper, we propose Momoka-RAG, which adopts a semantic completion-oriented approach to proactively explore the relationships between chunks in documents, thereby better addressing retrieval challenges in long-document scenarios. Momoka-RAG consists of two core components: Momoka-Map and Momoka-Trail Retriever. In Momoka-Map, we innovatively regard chunks rather than retrieval actions as nodes and utilize MCTS to discover the relationships between chunks, thereby constructing an information pathway model. In Momoka-Trail Retriever, we leverage the information pathway model obtained from the previous component to further expand the candidate pool, enabling more precise filtering of chunks relevant to user queries.

We conduct comparative experiments and generation tasks on datasets including Dragonball, SQUAD, NFCORPUS, SCI-DOCS, HotpotQA, and TriviaQA. The experimental results demonstrate that Momoka-RAG not only strikes a balance between accuracy and completeness when tackling long-document retrieval tasks but also efficiently prioritizes the most relevant chunks. Furthermore, the Momoka-RAG framework itself does not rely on specific embedding models or pre-trained LLM, nor does it involve additional training, making it applicable to a wide range of scenarios.

Limitations

During our research, we have identified several limitations of the proposed framework:

(1)Although Momoka-RAG advances beyond the passive and mechanical limitations of existing frameworks in reconstructing knowledge structures, it still cannot dynamically determine which seman-

tic completion path to take based on the user's specific query. The fundamental issue is that incorporating the user's query into the path construction process necessitates postponing the MCTS procedure until the retrieval stage. This would introduce significant latency due to real-time path generation, severely impacting system responsiveness. Conversely, executing MCTS before retrieval to avoid such delays makes it impossible to leverage the user's query to guide path exploration. Therefore, achieving query-adaptive dynamic path selection while avoiding the high computational overhead of MCTS remains a critical challenge. In future work, we will explore this trade-off in depth and strive to balance efficiency with dynamic semantic retrieval.

(2)Although we have shifted the substantial time required for the MCTS process to the indexing phase, the retrieval stage still necessitates invoking LLMs for filtering, resulting in higher latency compared to other frameworks. Furthermore, the quality of the MCTS procedure in Momoka-Map is closely tied to the number of iterations — more iterations lead to a higher-quality information pathway model, but also entail greater consumption of time and computational resources.

(3)Momoka-RAG employs MCTS to explore the relationships between chunks; however, these chunks are confined within a single document rather than spanning across multiple documents. In other words, our framework does not exhibit significant advantages when confronted with cross-document retrieval tasks. If all chunks from different documents are incorporated into the MCTS process, it would require an exceedingly large number of iterations, consuming substantial time and computational resources to establish the information pathway model. Furthermore, in such a scenario, positional information such as sentence indices and paragraph numbering would become ineffective for determining relationships between chunks across documents.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models. *Preprint*, arXiv:2310.05189.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval.

Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43.

Harrison Chase. 2024. Langchain: A framework for developing applications powered by language models. Accessed: 2024-12-13.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2309.07597.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense x retrieval: What retrieval granularity should we use? *Preprint*, arXiv:2312.06648.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *Preprint*, arXiv:2004.07180.

T. Cover and P. Hart. 2006. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, 13(1):21–27.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. Lightrag: Simple and fast retrieval-augmented generation. *Preprint*, arXiv:2410.05779.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. Hipporag: Neurobiologically inspired long-term memory for large language models. *Preprint*, arXiv:2405.14831.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. 2024. Late chunking: Contextual chunk embeddings using long-context embedding models. *Preprint*, arXiv:2409.04701.

Yunhai Hu, Yilun Zhao, Chen Zhao, and Arman Cohan. 2025. Mcts-rag: Enhancing retrieval-augmented generation with monte carlo tree search. *Preprint*, arXiv:2503.20757.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Levente Kocsis and Csaba Szepesvari. 2006. Bandit based monte-carlo planning. In *European Conference on Machine Learning*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. *Preprint*, arXiv:2003.08505.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *Preprint*, arXiv:2401.18059.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Meta-chunking: Learning efficient text segmentation via logical perception. *Preprint*, arXiv:2410.12788.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Rageval: Scenario specific rag evaluation dataset generation framework. *Preprint*, arXiv:2408.01262.

A Detailed Prompt in Momoka-RAG

In Figure 3 and 4, we have presented the prompt details of the Momoka-RAG.

B Case Study

In this section, we present case studies of various frameworks from the generation experiments in Tables 5, 6, 7, 8 and 9. In the responses generated by LightRAG, there is no mention of content related to "renewable energy" or "green finance" investments, nor of employee volunteer services. This indicates that the node associated with the "2020 Report" in its graph structure is not connected to nodes corresponding to the aforementioned content, and the links between nodes only capture local relationships. In contrast, the PPL and MSP methods of Meta-Chunking do not reorganize the knowledge structure. Although their semantically coherent text chunks may partially cover ground truths that are textually related, other cross-segmental and parallel content cannot be retrieved, resulting in suboptimal response quality.

Momoka-RAG, however, adopts a proactive exploration approach by retrieving semantically related chunks (in this case study, chunks with parallel relationships) together, enabling it to answer questions more comprehensively and effectively.

```
You are an information verification expert. An article has been divided into multiple
chunks, and you need to assess whether the new chunk semantically completes the
initial chunk—such as completing the subject, establishing a causal link, or adding a
parallel concept under the same theme.

If it does provide semantic completion, return 1; otherwise, return 0. Do not
output anything else.

Initial Chunk: {initial_chunk.text}
New Chunk: {new_chunk.text}
Output format: 1/0
```

Figure 3: Detailed prompt in Momoka-Map.

```
You are an expert in information comparison. Now, you need to determine whether
the **Chunks** is relevant to the **question**. If it is relevant, please output "1";
if not, please output "0".
If the **Chunks** have no relation to the **question**, output 0; if the
**Chunks** are related to the **question**, please output 1.

Chunks: {item}
Question: {question}
Output: 1 or 0
```

Figure 4: Detailed prompt in Momoka-Trail Retriever.

C Detailed Information of Datasets

The Dragonball dataset consists entirely of fictional information with no connection to real-world data. The SQUAD corpus is primarily sourced from Wikipedia articles. The medical documents in the NFCORPUS dataset are mainly from PubMed. The SCI-DOCS corpus includes scientific literature in fields such as computer science and physics. The HotpotQA is a large-scale multi-hop question-answering dataset based on Wikipedia. The TriviaQA is a large-scale dataset with its questions mostly derived from trivia questions found on the internet.

The details of each dataset can be found in Figures 10, 11, 12, 13, 14 and 15. The average chunk length for each dataset is shown in Table 5.

D Detailed Information of Comparative Experiments

In all experimental results tables, abbreviations were used for some framework names and metrics: PPL, MSP, DXR and Time_Ave refer to Meta-Chunking-PPL, Meta-Chunking-MSP, Dense X Retrieval and average time latency, respectively. Table 6 presents the complete experimental results of the metrics calculated according to formula 6. Tables 7, 8, 9, 10, 11 and 12 show the more detailed experimental results for each dataset.

Dataset	Average Chunk Length
Dragonball	22
SQUAD	63
NFCORPUS	25
SCI-DOCS	68
HotpotQA	53
TriviaQA	57

Table 5: Average chunk length of each dataset.

E Detailed Information of Generation Quality Metrics in Dragonball Datasets

In terms of generation quality, we employ the LLM prompt template used for automatic evaluation in the Dragonball dataset, along with the corresponding generation quality metrics, which include:

Relevant indicates that the information contained in the generated answer is core-relevant and consistent with key points in the standard answer.

Irrelevant indicates that the generated answer does not cover key points from the standard answer.

Wrong indicates that the generated answer cov-

Question:
Based on Acme Financial Services' 2020 Corporate Governance Report, summarize the sustainability and social responsibility initiatives taken by the company.

"ground_truth_keypoints":
[
"1. Acme Financial Services' 2020 Corporate Governance Report highlights sustainability and social responsibility initiatives.",

"2. The company invested in renewable energy and engaged in green financing.",

"3. Acme Financial Services promoted employee volunteering.",

"4. These initiatives enhanced the company's public image and improved its market competitiveness.",

"5. The company is dedicated to integrating sustainability and social responsibility into its business strategy."
]

Figure 5: Case study on generation quality experiment.

Momoka-RAG

Question:
Based on Acme Financial Services' 2020 Corporate Governance Report, summarize the sustainability and social responsibility initiatives taken by the company.

"generated_answer": "Based on Acme Financial Services' 2020 Corporate Governance Report, the company demonstrated a strong commitment to sustainability and social responsibility, particularly highlighted in July 2020. Key initiatives included:\n\n1. **Investment in Renewable Energy**: Acme Financial Services invested in renewable energy projects, aligning its operations with environmentally sustainable practices.\n \n2. **Green Financing**: The company engaged in green financing, supporting projects and activities that contribute to environmental sustainability.\n\n3. **Employee Volunteering**: Acme encouraged employee volunteering, fostering a culture of social responsibility and community engagement among its workforce.\n\nAdditionally, in March 2020, the company revised its corporate governance policy to enhance transparency, accountability, and stakeholder engagement, further reinforcing its dedication to responsible business practices. These efforts have positioned Acme Financial Services as a responsible and trustworthy financial institution within the industry. \n\nOverall, Acme Financial Services has shown a clear focus on integrating sustainability and social responsibility into its core operations, contributing positively to both the environment and society.",

"evaluation_result": {"completeness": 1.0, "hallucination": 0.0, "irrelevance": 0.0}

Figure 6: Case study of Momoka-RAG.

LightRAG

Question:
Based on Acme Financial Services' 2020 Corporate Governance Report, summarize the sustainability and social responsibility initiatives taken by the company.

"generated_answer": "### Sustainability and Social Responsibility Initiatives by Acme Financial Services\n\nAcme Financial Services has demonstrated a strong commitment to sustainability and social responsibility, particularly evident in its actions during 2020. In July 2020, the company highlighted its dedication to these principles, aligning with broader corporate values aimed at environmental impact and societal contribution. The organization revised its corporate governance policy in March 2020 to enhance transparency, accountability, and stakeholder engagement, which supported its sustainability efforts. Additionally, Acme Financial Services expanded its service offerings by introducing a new division specializing in insurance services in March 2020, diversifying its portfolio while maintaining a focus on sustainable practices. The company also implemented stricter internal control measures during the same period, positively impacting its reputation and reinforcing its commitment to ethical business practices. Furthermore, Acme Financial Services launched a new investment product targeting high-net-worth individuals in November 2020, reflecting its strategy to integrate sustainability within its financial offerings.\n\n### References\n1. [DC] unknown_source\n2. [DC] unknown_source\n3. [KG] unknown_source\n4. [KG] unknown_source\n5. [KG] unknown_source",

"evaluation_result": {"completeness": 0.4, "hallucination": 0.0, "irrelevance": 0.6}

Figure 7: Case study of LightRAG.

PPL

Question:
Based on Acme Financial Services' 2020 Corporate Governance Report, summarize the sustainability and social responsibility initiatives taken by the company.

"generated_answer": "Based on Acme Financial Services' 2020 Corporate Governance Report, the company demonstrated a strong commitment to sustainability and social responsibility through several key initiatives:\n\n1. **Commitment to Sustainability and Social Responsibility (July 2020):** \n Acme Financial Services explicitly highlighted its dedication to sustainability and social responsibility in July 2020. While specific details of the sustainability initiatives are not provided, this commitment aligns with the company's broader focus on ethical practices and stakeholder engagement.\n\n2. **Enhanced Corporate Governance Policy (March 2020):** \n The company revised its corporate governance policy to improve transparency, accountability, and stakeholder engagement. This revision ensured adherence to best practices and reflected the company's dedication to meeting stakeholder expectations, which is a critical aspect of social responsibility.\n\n3. **Addressing Ethics and Integrity Incidents:** \n Following a significant ethics and integrity incident uncovered in February 2020, Acme Financial Services took decisive actions to restore trust and strengthen its governance framework. These included conducting internal investigations, terminating employees involved in fraudulent activities, and implementing stricter internal control measures. Such steps underscore the company's commitment to ethical business practices, which are foundational to social responsibility.\n\n4. **Engagement with External Audit Firms (April 2020):** \n To further enhance transparency and accountability, the company engaged an external audit firm for independent verification. This move reinforced trust among investors and stakeholders, demonstrating the company's proactive approach to maintaining high governance standards.\n\n5. **Staff Training Programs (May 2020):** \n Acme Financial Services enhanced its staff training programs to promote ethics and integrity within the organization. These programs aimed to improve employee knowledge and awareness of ethical practices, fostering a culture of responsibility and accountability.\n\n6. **Focus on Long-Term Growth Drivers:** \n Beyond immediate governance reforms, the company emphasized long-term strategies such as cost control, technological advancements, and risk management. These efforts not only support sustainable growth but also align with responsible business practices that benefit both the company and society.\n\nIn summary, Acme Financial Services' sustainability and social responsibility initiatives in 2020 were centered around strengthening corporate governance, promoting ethical practices, enhancing transparency, and engaging stakeholders. These efforts collectively positioned the company as a responsible and trustworthy financial institution committed to positive societal impact.\n\n**Final Answer:**\n**{Acme Financial Services demonstrated its commitment to sustainability and social responsibility in 2020 by revising its corporate governance policies, addressing ethics incidents, engaging external auditors, enhancing staff training programs, and focusing on long-term growth drivers like cost control and risk management. These initiatives emphasized transparency, accountability, and stakeholder engagement, reinforcing the company's role as a responsible financial institution.}**",

"evaluation_result": {"completeness": 0.4, "hallucination": 0.0, "irrelevance": 0.6}

Figure 8: Case study of PPL.

MSP

Question:
Based on Acme Financial Services' 2020 Corporate Governance Report, summarize the sustainability and social responsibility initiatives taken by the company.

"generated_answer": "Based on the information provided in Acme Financial Services' 2020 Corporate Governance Report, the company took several notable sustainability and social responsibility initiatives:\n\n1. **Commitment to Sustainability and Social Responsibility (July 2020)**: Acme Financial Services explicitly demonstrated its commitment to sustainability and social responsibility in July 2020. This initiative was part of the company's broader effort to align its operations with environmentally friendly practices and contribute positively to society.\n\n2. **Revised Corporate Governance Policy (March 2020)**: In March 2020, the company revised its corporate governance policy to enhance transparency, accountability, and stakeholder engagement. These changes ensured that Acme adhered to best practices in governance while meeting the expectations of its stakeholders, which indirectly supported its sustainability goals by fostering trust and ethical practices.\n\n3. **Enhanced Staff Training Programs (May 2020)**: In May 2020, Acme Financial Services improved its staff training programs to promote ethics and integrity within the organization. By increasing employee awareness of ethical practices, the company laid a foundation for responsible decision-making, which is critical for sustainability and social responsibility.\n\n4. **Response to Ethics Incident (February 2020)**: Although not directly related to sustainability, the ethics and integrity incident in February 2020 likely prompted the company to strengthen its focus on responsible governance and corporate citizenship as part of its recovery efforts.\n\nWhile these initiatives highlight Acme Financial Services' progress in 2020, it is worth noting that the company continued to build on these efforts in subsequent years, as evidenced by further advancements in sustainability and social responsibility reported in June 2021.\n\nIn summary, Acme Financial Services' 2020 initiatives focused on enhancing governance practices, promoting ethical behavior, and committing to sustainability and social responsibility, all of which contributed to the company's long-term reputation and stakeholder trust. \n\n**Final Answer**: \n$$\n\\boxed{\\text{Acme Financial Services demonstrated its commitment to sustainability and social responsibility in 2020 through enhanced governance policies, staff training on ethics, and transparent practices.}}\n$$",

"evaluation_result": {"completeness": 0.4, "hallucination": 0.0, "irrelevance": 0.6}

Figure 9: Case study of MSP.

ers key points from the standard answer, but the information is incorrect or contradictory to the standard answer points.

The specific prompt template can be found in the code repository of the RAGEval project [2].

F Detailed Information of Statistical Validation Experiment

This section presents the detailed results of the statistical significance experiments, which are provided in Tables 13, 14, 15, 16, 17 and 18.

G Ablation Studies

In the Momoka-RAG framework, since the functional implementations of Momoka-Map and Momoka-Trail Retriever cannot be separated into individual components for ablation studies, this section only compares the performance differences between Naive RAG and Momoka-RAG using the same chunk sequences. For details, please refer to Tables 19, 20, 21, 22, 23, 24 and 25.

[2] https://github.com/OpenBMB/RAGEval

Methods	Momoka-RAG	Late-Chunking	RAPTOR	PPL	MSP	DXR
Dragonball						
Recall	82.44	2.23	64.19	127.13	87.03	3.29
MRR	2.28	0.04	1.37	2.20	1.85	0.11
Precision	218.21	3.51	101.08	133.86	111.39	6.14
EIR	83.30	127.43	158.71	92.20	130.77	166.18
F1	1.18	0.03	0.71	1.16	0.88	0.03
SQUAD						
Recall	289.67	70.67	/	252.79	260.90	191.06
MRR	2.88	0.45	/	2.47	2.49	1.68
Precision	279.34	34.58	/	133.26	134.77	96.24
F1	2.84	0.43	/	1.61	1.64	1.19
NFCORPUS						
Recall	279.21	12.95	285.88	285.10	283.92	283.92
MRR	2.79	0.11	2.70	2.79	2.75	2.74
Precision	277.48	4.79	261.30	243.53	251.84	264.79
F1	2.78	0.07	2.73	2.62	2.66	2.73
SCI-DOCS						
Recall	280.05	1.71	291.93	286.55	289.90	287.00
MRR	2.79	0.03	2.85	2.83	2.86	2.83
Precision	279.25	1.02	275.1	264.17	266.64	270.14
F1	2.80	0.02	2.83	2.74	2.78	2.79
HotpotQA						
Recall	286.07	11.40	/	84.81	281.01	237.35
MRR	2.85	0.12	/	0.82	2.80	2.30
Precision	283.24	7.81	/	69.07	237.60	118.78
F1	2.84	0.10	/	0.76	2.56	1.42
TriviaQA						
Recall	81.47	46.72	/	14.29	57.92	37.83
MRR	0.81	0.47	/	0.01	0.15	0.38
Precision	81.47	46.72	/	14.29	57.92	37.83
F1	0.81	0.47	/	0.14	0.58	0.38

Table 6: Detail information of comparative experiments.

Methods	Momoka-RAG	Late-Chunking	RAPTOR	PPL	MSP	DXR	LightRAG
Top-1							
Recall	20.54	0.52	14.85	27.13	18.49	0.75	6.61
MRR	0.72	0.02	0.55	0.69	0.57	0.03	0.24
Precision	72.02	1.83	54.53	68.97	56.56	3.26	23.70
EIR	49.77	71.30	92.75	51.52	74.03	100.00	90.35
F1	0.32	0.01	0.23	0.39	0.28	0.01	0.10
Time_Ave	2.28	0.04	0.12	0.03	0.03	0.03	/
Top-3							
Recall	27.34	0.75	22.52	45.22	30.56	1.18	10.92
MRR	0.79	0.01	0.45	0.75	0.63	0.04	0.22
Precision	76.40	0.95	27.08	37.67	31.43	1.70	12.72
EIR	19.51	32.04	40.08	24.03	34.01	40.70	31.64
F1	0.40	0.01	0.25	0.41	0.31	0.01	0.12
Time_Ave	6.73	0.04	0.30	0.03	0.03	0.03	/
Top-5							
Recall	34.56	0.96	26.82	54.78	37.98	1.36	14.00
MRR	0.77	0.01	0.37	0.76	0.65	0.04	0.19
Precision	69.79	0.73	19.47	27.22	23.40	1.18	9.81
EIR	14.02	24.09	25.88	16.65	22.73	25.48	20.47
F1	0.46	0.01	0.23	0.36	0.29	0.01	0.12
Time_Ave	12.03	0.04	0.32	0.03	0.03	0.03	/

Table 7: Detail information of comparative experiments of Dragonball.

Methods	Momoka-RAG	Late-Chunking	RAPTOR	PPL	MSP	DXR
			Top-1			
Recall	**95.53**	**16.48**	**/**	**80.73**	**79.89**	**50.84**
MRR	**0.96**	**0.16**	**/**	**0.81**	**0.80**	**0.51**
Precision	**95.53**	**16.48**	**/**	**80.73**	**79.89**	**50.84**
F1	**0.96**	**0.16**	**/**	**0.81**	**0.8**	**0.51**
Time_Ave	**3.64**	**0.02**	**/**	**0.03**	**0.03**	**0.03**
			Top-3			
Recall	**96.65**	**25.70**	**/**	**85.75**	**87.71**	**66.76**
MRR	**0.96**	**0.16**	**/**	**0.83**	**0.84**	**0.58**
Precision	**92.36**	**10.61**	**/**	**32.31**	**32.87**	**26.35**
F1	**0.94**	**0.15**	**/**	**0.47**	**0.48**	**0.38**
Time_Ave	**11.32**	**0.02**	**/**	**0.03**	**0.03**	**0.03**
			Top-5			
Recall	**97.49**	**28.49**	**/**	**86.31**	**93.30**	**73.46**
MRR	**0.96**	**0.13**	**/**	**0.83**	**0.85**	**0.59**
Precision	**91.45**	**7.49**	**/**	**20.22**	**22.01**	**19.05**
F1	**0.94**	**0.12**	**/**	**0.33**	**0.36**	**0.30**
Time_Ave	**19.26**	**0.03**	**/**	**0.03**	**0.03**	**0.03**

Table 8: Detail information of comparative experiments of SQUAD.

Methods	Momoka-RAG	Late-Chunking	RAPTOR	PPL	MSP	DXR
			Top-1			
Recall	**91.76**	**2.75**	**90.59**	**92.16**	**89.41**	**89.41**
MRR	**0.92**	**0.03**	**0.91**	**0.92**	**0.89**	**0.89**
Precision	**91.76**	**2.75**	**90.59**	**92.16**	**89.41**	**89.41**
F1	**0.92**	**0.03**	**0.91**	**0.92**	**0.89**	**0.89**
Time_Ave	**5.13**	**0.03**	**0.18**	**0.03**	**0.03**	**0.03**
			Top-3			
Recall	**92.55**	**5.10**	**96.86**	**94.90**	**96.47**	**96.47**
MRR	**0.92**	**0.04**	**0.91**	**0.93**	**0.93**	**0.92**
Precision	**91.76**	**1.02**	**87.43**	**81.57**	**83.92**	**88.63**
F1	**0.92**	**0.02**	**0.92**	**0.88**	**0.90**	**0.92**
Time_Ave	**17.14**	**0.03**	**0.31**	**0.03**	**0.03**	**0.03**
			Top-5			
Recall	**94.90**	**5.10**	**98.43**	**98.04**	**98.04**	**98.04**
MRR	**0.95**	**0.04**	**0.88**	**0.94**	**0.93**	**0.93**
Precision	**93.96**	**1.02**	**83.28**	**69.8**	**78.51**	**86.75**
F1	**0.94**	**0.02**	**0.90**	**0.82**	**0.87**	**0.92**
Time_Ave	**29.27**	**0.03**	**0.33**	**0.03**	**0.03**	**0.03**

Table 9: Detail information of comparative experiments of NFCORPUS.

Methods	Momoka-RAG	Late-Chunking	RAPTOR	PPL	MSP	DXR
			Top-1			
Recall	**92.60**	**0.37**	**95.07**	**93.27**	**93.72**	**92.83**
MRR	**0.93**	**0.01**	**0.95**	**0.93**	**0.94**	**0.93**
Precision	**92.60**	**0.67**	**95.07**	**93.27**	**93.72**	**92.83**
F1	**0.93**	**0.01**	**0.95**	**0.93**	**0.94**	**0.93**
Time_Ave	**6.01**	**0.09**	**0.17**	**0.02**	**0.03**	**0.03**
			Top-3			
Recall	**93.50**	**0.67**	**97.98**	**96.19**	**97.53**	**96.64**
MRR	**0.93**	**0.01**	**0.96**	**0.95**	**0.96**	**0.95**
Precision	**93.42**	**0.22**	**91.62**	**88.12**	**88.12**	**89.01**
F1	**0.93**	**0.003**	**0.95**	**0.92**	**0.93**	**0.93**
Time_Ave	**19.35**	**0.08**	**0.36**	**0.03**	**0.03**	**0.03**
			Top-5			
Recall	**93.95**	**0.67**	**98.88**	**97.09**	**98.65**	**97.53**
MRR	**0.93**	**0.01**	**0.94**	**0.95**	**0.96**	**0.95**
Precision	**93.23**	**0.13**	**88.41**	**82.78**	**84.8**	**88.3**
F1	**0.94**	**0.002**	**0.93**	**0.89**	**0.91**	**0.93**
Time_Ave	**33.11**	**0.08**	**1.56**	**0.03**	**0.03**	**0.03**

Table 10: Detail information of comparative experiments of SCI-DOCS.

Methods	Momoka-RAG	Late-Chunking	RAPTOR	PPL	MSP	DXR
			Top-1			
Recall	**94.30**	**3.80**	**/**	**24.68**	**93.04**	**75.32**
MRR	**0.94**	**0.04**	**/**	**0.25**	**0.93**	**0.75**
Precision	**94.3**	**0.04**	**/**	**24.68**	**93.04**	**75.32**
F1	**0.94**	**3.80**	**/**	**0.25**	**0.93**	**0.75**
Time_Ave	**4.97**	**0.05**	**/**	**0.03**	**0.03**	**0.03**
			Top-3			
Recall	**95.57**	**3.80**	**/**	**29.75**	**93.67**	**80.38**
MRR	**0.95**	**0.04**	**/**	**0.29**	**0.93**	**0.77**
Precision	**94.51**	**2.11**	**/**	**24.26**	**78.48**	**27.00**
F1	**0.95**	**0.03**	**/**	**0.27**	**0.85**	**0.40**
Time_Ave	**16.06**	**0.05**	**/**	**0.03**	**0.03**	**0.04**
			Top-5			
Recall	**96.20**	**3.80**	**/**	**30.38**	**94.30**	**81.65**
MRR	**0.96**	**0.04**	**/**	**0.28**	**0.94**	**0.78**
Precision	**94.43**	**1.90**	**/**	**20.13**	**66.08**	**16.46**
F1	**0.95**	**0.03**	**/**	**0.24**	**0.78**	**0.27**
Time_Ave	**28.52**	**0.05**	**/**	**0.03**	**0.03**	**0.03**

Table 11: Detail information of comparative experiments of HotpotQA.

Methods	Momoka-RAG	Late-Chunking	RAPTOR	PPL	MSP	DXR
			Top-1			
Recall	**81.47**	**46.72**	**/**	**14.29**	**57.92**	**37.83**
MRR	**0.81**	**0.47**	**/**	**0.01**	**0.15**	**0.38**
Precision	**81.47**	**46.72**	**/**	**14.29**	**57.92**	**37.83**
F1	**0.81**	**0.47**	**/**	**0.14**	**0.58**	**0.38**
Time_Ave	**3.29**	**0.05**	**/**	**0.03**	**0.03**	**0.03**

Table 12: Detail information of comparative experiments of TriviaQA.

Methods	RAPTOR	PPL	MSP	DXR	LightRAG
Top-1					
Recall	/	/	/	/	3.33E-05★
MRR	3.33E-05★	5.40E-02	3.33E-05★	3.33E-05★	3.33E-05★
Precision	3.33E-05★	5.55E-02	3.33E-05★	3.33E-05★	3.33E-05★
F1	/	/	/	/	3.33E-05★
Top-3					
Recall	/	/	/	/	3.33E-05★
MRR	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★
Precision	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★
F1	/	/	/	/	3.33E-05★
Top-5					
Recall	/	/	/	/	3.33E-05★
MRR	3.3E-05	8.34E-01	3.33E-05★	3.33E-05★	3.33E-05★
Precision	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★

Table 13: Statistical validation experiments on Dragonball.

Methods	PPL	MSP	DXR
Top-1			
Recall	3.33E-05★	3.33E-05★	3.33E-05★
MRR	3.33E-05★	3.33E-05★	3.33E-05★
Precision	3.33E-05★	3.33E-05★	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★
Top-3			
Recall	3.33E-05★	3.33E-05★	3.33E-05★
MRR	3.33E-05★	3.33E-05★	3.33E-05★
Precision	3.33E-05★	3.33E-05★	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★
Top-5			
Recall	3.33E-05★	3.33E-05★	3.33E-05★
MRR	3.33E-05★	3.33E-05★	3.33E-05★
Precision	3.33E-05★	1.04E-03	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★

Table 14: Statistical validation experiments on SQUAD.

Methods	RAPTOR	PPL	MSP	DXR
Top-3				
Precision	6.73E-03★	3.33E-05★	3.33E-05★	8.13E-02
F1	3.31E-01	5.67E-04★	1.39E-02★	5.73E-01
Top-5				
MRR	2.67E-04★	5.21E-01	3.04E-01	1.98E-01
Precision	3.33E-05★	3.33E-05★	3.33E-05★	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★	3.37E-03

Table 15: Statistical validation experiments on NFCORPUS.

Methods	RAPTOR	PPL	MSP	DXR
Top-3				
Precision	1.02E-01	6.67E-05★	3.33E-05★	6.00E-04★
Top-5				
Precision	2.00E-04	3.33E-05★	3.33E-05★	2.33E-04★
F1	1.61E-01	3.33E-05★	5.33E-04★	9.03E-02

Table 16: Statistical validation experiments on SCI-DOCS.

Methods	PPL	MSP	DXR
Top-1			
Recall	3.33E-05★	7.57E-01	3.33E-05★
MRR	3.33E-05★	7.56E-01	3.33E-05★
Precision	3.33E-05★	7.50E-01	3.33E-05★
F1	3.33E-05★	7.48E-01	3.33E-05★
Top-3			
Recall	3.33E-05★	5.47E-01	3.33E-05★
MRR	3.33E-05★	4.51E-01	3.33E-05★
Precision	3.33E-05★	3.33E-05★	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★
Top-5			
Recall	3.33E-05★	5.50E-01	3.33E-05★
MRR	3.33E-05★	3.06E-01	3.33E-05★
Precision	3.33E-05★	3.33E-05★	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★

Table 17: Statistical validation experiments on HotpotQA.

Methods	PPL	MSP	DXR
Top-1			
Recall	3.33E-05★	3.33E-05★	3.33E-05★
MRR	3.33E-05★	3.33E-05★	3.33E-05★
Precision	3.33E-05★	3.33E-05★	3.33E-05★
F1	3.33E-05★	3.33E-05★	3.33E-05★

Table 18: Statistical validation experiments on TriviaQA.

{"domain": "Finance",
"language": "en",
"query": {"query_id": 2200,
"query_type": "Summary Question",
"content": "Based on Grand Adventures Tourism Ltd.'s 2021 report, summarize the financial and ethical challenges the company faced and the measures taken to address them."},

"ground_truth": {"doc_ids": [50],

"content": "In 2021, Grand Adventures Tourism Ltd. faced several financial and ethical challenges. In January, the company encountered significant ethical or integrity violations, including fraud and conflicts of interest. An internal audit in May revealed financial improprieties and suspicious transactions, raising concerns about the accuracy of financial reports. In response, the board of directors launched a formal investigation in June, leading to the suspension of senior management in July. The company announced the need to restate its financial statements in August due to identified errors and misstatements. To address these issues, a reputable forensic accounting firm was hired in September to conduct a detailed investigation. These measures demonstrated the company's commitment to uncovering the truth, ensuring transparency, and holding individuals accountable for unethical behavior.",

"references": ["One of the most notable events that occurred in January 2021 was the emergence of ethics and integrity incidents within the company.", "These incidents involved significant violations, such as fraud, corruption, and conflicts of interest.", "To address these issues, Grand Adventures Tourism Ltd. took several measures, including launching an internal audit in May 2021.", "The audit revealed financial improprieties and suspicious transactions, raising concerns about the accuracy and integrity of the company's financial reports.", "In response to the internal audit findings and alleged ethics and integrity incidents, the board of directors initiated a formal investigation in June 2021.", "This investigation aimed to uncover the truth behind the allegations and demonstrate the company's commitment to addressing the issues at hand.", "As a result of the investigation, senior executives implicated in the internal audit findings and ethics and integrity incidents were placed on suspension in July 2021, pending the outcome of the investigation.", "This action sent a strong message that Grand Adventures Tourism Ltd. would not tolerate unethical behavior and would hold individuals accountable.", "Furthermore, in August 2021, the company announced the need to restate its financial statements due to identified errors and misstatements.", "This restatement raised concerns about the accuracy and reliability of previously reported financial information, potentially damaging investor trust.", "To ensure a thorough investigation into the financial improprieties, Grand Adventures Tourism Ltd. hired a reputable forensic accounting firm in September 2021."],
.....}
{"domain": "Finance",
"language": "en",
"query": {"query_id": 2252,
"query_type": "Multi-hop Reasoning Question",
"content": "How did the corporate governance policy revision in January 2018, including the appointment of independent board members, the implementation of a whistle-blower program, and the introduction of a board evaluation process, enhance stakeholder confidence in CleanCo Housekeeping Services?"},

"ground_truth": {"doc_ids": [47],

"content": "The corporate governance policy revision in January 2018 included several key measures: the appointment of independent board members increased the board's diversity and expertise, enhancing transparency in decision-making; the implementation of a whistle-blower program provided a mechanism to detect and address unethical behavior, strengthening corporate governance practices; and the introduction of a board evaluation process identified areas for improvement in governance. Collectively, these measures enhanced transparency, accountability, and stakeholder engagement, thereby boosting stakeholder confidence in CleanCo Housekeeping Services.",

"references": ["Firstly, in January 2018, the company revised its corporate governance policies to enhance transparency, accountability, and stakeholder engagement.", "This revision included the implementation of regular board evaluations, strengthened codes of ethics, and increased shareholder communication programs.", "These changes aimed to improve corporate governance transparency and accountability, ultimately boosting stakeholder confidence.", "As part of the policy revision, CleanCo Housekeeping Services appointed three independent board members in March 2018.", "These board members brought diverse backgrounds and expertise to the company, strengthening the board's independence and enhancing transparency in decision-making processes.", "The appointment of independent board members increased the board's diversity and expertise, leading to improved governance and decision-making.", "In June 2018, CleanCo Housekeeping Services introduced a whistle-blower program to allow employees, clients, and other stakeholders to report any unethical behavior, fraud, or violations of corporate governance policies anonymously and with protection from retaliation.", "This program created a mechanism to detect and address unethical behavior, further strengthening the company's corporate governance practices.", "Another important event in September 2018 was the implementation of a comprehensive board evaluation process.", "This process aimed to assess the effectiveness of the board, its committees, and individual directors.", "The evaluation included self-assessment and external evaluation to identify areas for improvement in corporate governance practices.", "This initiative played a crucial role in identifying weaknesses and areas for improvement, facilitating better decision-making and accountability within the company."],
..... }

Figure 10: Detailed informations of Dragonball dataset.

{'id': '572651f9f1498d1400e8dbf2',
'title': 'European_Union_law',
'context': 'While the Commission has a monopoly on initiating legislation, the European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process. According to the Treaty on European Union articles 9 and 10, the EU observes "the principle of equality of its citizens" and is meant to be founded on "representative democracy". In practice, equality and democracy are deficient because the elected representatives in the Parliament cannot initiate legislation against the Commission\'s wishes, citizens of smallest countries have ten times the voting weight in Parliament as citizens of the largest countries, and "qualified majorities" or consensus of the Council are required to legislate. The justification for this "democratic deficit" under the Treaties is usually thought to be that completion integration of the European economy and political institutions required the technical coordination of experts, while popular understanding of the EU developed and nationalist sentiments declined post-war. Over time, this has meant the Parliament gradually assumed more voice: from being an unelected assembly, to its first direct elections in 1979, to having increasingly more rights in the legislative process. Citizens\' rights are therefore limited compared to the democratic polities within all European member states: under TEU article 11 citizens and associations have the rights such as publicising their views and submit an initiative that must be considered by the Commission with one million signatures. TFEU article 227 contains a further right for citizens to petition the Parliament on issues which affect them. Parliament elections, take place every five years, and votes for Members of the European Parliament in member states must be organised by proportional representation or a single transferable vote. There are 750 MEPs and their numbers are "degressively proportional" according to member state size. This means - although the Council is meant to be the body representing member states - in the Parliament citizens of smaller member states have more voice than citizens in larger member states. MEPs divide, as they do in national Parliaments, along political party lines: the conservative European People\'s Party is currently the largest, and the Party of European Socialists leads the opposition. Parties do not receive public funds from the EU, as the Court of Justice held in Parti écologiste "Les Verts" v Parliament that this was entirely an issue to be regulated by the member states. The Parliament\'s powers include calling inquiries into maladministration or appoint an Ombudsman pending any court proceedings. It can require the Commission respond to questions and by a two-thirds majority can censure the whole Commission (as happened to the Santer Commission in 1999). ...',

'question': 'What two bodies must the Parliament go through first to pass legislation?',
'answers': {'
text': ['the Commission and Council', 'the Commission and Council', 'the Commission and Council', 'the European Parliament and the Council of the European Union'],
'answer_start': [3090, 3090, 3090, 63]}}

Figure 11: Detailed informations of SQUAD dataset.

{"_id": "MED-946",
"title": "Bulking agents, antispasmodics and antidepressants for the treatment of irritable bowel syndrome.",
"text": "BACKGROUND: Irritable bowel syndrome (IBS) is a common chronic gastrointestinal disorder. The role of pharmacotherapy for IBS is limited and focused mainly on symptom control. OBJECTIVES: The objective of this systematic review was to evaluate the efficacy of bulking agents, antispasmodics and antidepressants for the treatment of irritable bowel syndrome. SEARCH STRATEGY: Computer assisted structured searches of MEDLINE, EMBASE, The Cochrane library, CINAHL and PsychInfo were conducted for the years 1966-2009. An updated search in April 2011 identified 10 studies which will be considered for inclusion in a future update of this review. SELECTION CRITERIA: Randomized controlled trials comparing bulking agents, antispasmodics or antidepressants with a placebo treatment in patients with irritable bowel syndrome aged over 12 years were considered for inclusion. Only studies published as full papers were included. Studies were not excluded on the basis of language. The primary outcome had to include improvement of abdominal pain, global assessment or symptom score. DATA COLLECTION AND ANALYSIS: Two authors independently extracted data from the selected studies. Risk Ratios (RR) and Standardized Mean Differences (SMD) with 95% confidence intervals (CI) were calculated. A proof of practice analysis was conducted including sub-group analyses for different types of bulking agents, spasmolytic agents or antidepressant medication. This was followed by a proof of principle analysis where only the studies with adequate allocation concealment were included. MAIN RESULTS: A total of 56 studies (3725 patients) were included in this review. These included 12 studies of bulking agents (621 patients), 29 of antispasmodics (2333 patients), and 15 of antidepressants (922 patients). The risk of bias was low for most items. However, selection bias is unclear for many of the included studies because the methods used for randomization and allocation concealment were not described. No beneficial effect for bulking agents over placebo was found for improvement of abdominal pain (4 studies; 186 patients; SMD 0.03; 95% CI -0.34 to 0.40; P = 0.87), global assessment (11 studies; 565 patients; RR 1.10; 95% CI 0.91 to 1.33; P = 0.32) or symptom score (3 studies; 126 patients SMD -0.00; 95% CI -0.43 to 0.43; P = 1.00). Subgroup analyses for insoluble and soluble fibres also showed no statistically significant benefit. Separate analysis of the studies with adequate concealment of allocation did not change these results. There was a beneficial effect for antispasmodics over placebo for improvement of abdominal pain (58% of antispasmodic patients improved compared to 46% of placebo; 13 studies; 1392 patients; RR 1.32; 95% CI 1.12 to 1.55; P < 0.001; NNT = 7), global assessment (57% of antispasmodic patients improved compared to 39% of placebo; 22 studies; 1983 patients; RR 1.49; 95% CI 1.25 to 1.77; P < 0.0001; NNT = 5) and symptom score (37% of antispasmodic patients improved compared to 22% of placebo; 4 studies; 586 patients; RR 1.86; 95% CI 1.26 to 2.76; P < 0.01; NNT = 3). ...",

"query": "why was bulking used for irritable bowel syndrome"}

Figure 12: Detailed informations of NFCORPUS dataset.

Methods	Naive	Momoka-RAG
Dragonball		
Recall	63.65	82.44
MRR	1.74	2.28
Precision	100.07	218.21
EIR	160.94	83.30
F1	0.70	1.18
SQUAD		
Recall	277.66	289.67
MRR	2.64	2.88
Precision	144.21	279.34
F1	1.75	2.84
NFCORPUS		
Recall	282.74	279.21
MRR	2.76	2.79
Precision	254.03	277.48
F1	2.68	2.78
SCI-DOCS		
Recall	281.16	280.05
MRR	2.77	2.79
Precision	261.18	279.25
F1	2.71	2.80
HotpotQA		
Recall	277.22	286.07
MRR	2.74	2.85
Precision	249.84	283.24
F1	2.63	2.84
TriviaQA		
Recall	64.86	81.47
MRR	0.11	0.81
Precision	64.86	81.47
F1	0.65	0.81

Table 19: **Ablation Studies:** In the table, **"Naive"** represents Naive RAG, which means embedding the embedding model directly into the vector space while keeping the chunk size consistent, and using cosine similarity for retrieval. In the TriviaQA dataset, due to the increase in Top-k, chunks containing sensitive information were retrieved, causing the LLM component used for filtering in Momoka-RAG to fail. Therefore, we only present ablation experiment results of Top-1.

Methods	Naive	Momoka-RAG
Top-1		
Recall	14.73	20.54
MRR	0.54	0.72
Precision	53.81	72.02
EIR	92.78	49.77
F1	0.23	0.32
Time_Ave	0.04	2.28
Top-3		
Recall	82.44	64.19
MRR	2.28	1.37
Precision	218.21	101.08
EIR	83.30	158.71
F1	1.18	0.71
Time_Ave	0.04	6.73
Top-5		
Recall	26.47	34.56
MRR	0.61	0.77
Precision	19.23	69.79
EIR	26.83	14.02
F1	0.22	0.46
Time_Ave	0.04	12.03

Table 20: Detail information of ablation studies of Dragonball.

Methods	Naive	Momoka-RAG
Top-1		
Recall	84.36	95.53
MRR	0.84	0.96
Precision	84.36	95.53
F1	0.84	0.96
Time_Ave	0.03	3.64
Top-3		
Recall	95.81	96.65
MRR	0.90	0.96
Precision	36.22	92.36
F1	0.53	0.94
Time_Ave	0.03	11.32
Top-5		
Recall	97.49	97.49
MRR	0.90	0.96
Precision	23.63	91.45
F1	0.38	0.94
Time_Ave	0.03	19.26

Table 21: Detail information of ablation studies of SQUAD.

Methods	Naive	Momoka-RAG
Top-1		
Recall	89.80	91.76
MRR	0.90	0.92
Precision	89.80	91.76
F1	0.90	0.92
Time_Ave	0.04	5.13
Top-3		
Recall	95.69	92.55
MRR	0.93	0.92
Precision	84.31	91.76
F1	0.90	0.92
Time_Ave	0.03	17.14
Top-5		
Recall	97.25	94.90
MRR	0.93	0.95
Precision	79.92	93.96
F1	0.88	0.94
Time_Ave	0.002	29.27

Table 22: Detail information of ablation studies of NFCORPUS.

Methods	Naive	Momoka-RAG
Top-1		
Recall	91.70	92.60
MRR	0.92	0.93
Precision	91.70	92.60
F1	0.92	0.93
Time_Ave	0.04	6.01
Top-3		
Recall	94.84	93.50
MRR	0.93	0.93
Precision	87.37	93.42
F1	0.91	0.93
Time_Ave	0.04	19.35
Top-5		
Recall	94.62	93.95
MRR	0.92	0.93
Precision	82.11	93.23
F1	0.88	0.94
Time_Ave	0.04	33.11

Table 23: Detail information of ablation studies of SCI-DOCS.

Methods	Naive	Momoka-RAG
Top-1		
Recall	90.51	94.30
MRR	0.91	0.94
Precision	90.51	94.30
F1	0.91	0.94
Time_Ave	0.03	4.97
Top-3		
Recall	91.77	95.57
MRR	0.91	0.95
Precision	82.49	94.51
F1	0.87	0.95
Time_Ave	0.04	16.06
Top-5		
Recall	94.94	96.20
MRR	0.92	0.96
Precision	76.84	94.43
F1	0.85	0.95
Time_Ave	0.04	28.52

Table 24: Detail information of ablation studies of HotpotQA.

Methods	Naive	Momoka-RAG
Top-1		
Recall	64.86	81.47
MRR	0.11	0.81
Precision	64.86	81.47
F1	0.65	0.81
Time_Ave	0.03	3.29

Table 25: Detail information of ablation studies of TriviaQA.

{"_id": "011d4ccb74f32f597df54ac8037a7903bd95038b",
"title": "The evolution of human skin coloration.",
"text": "Skin color is one of the most conspicuous ways in which humans vary and has been widely used to define human races. Here we present new evidence indicating that variations in skin color are adaptive, and are related to the regulation of ultraviolet (UV) radiation penetration in the integument and its direct and indirect effects on fitness. Using remotely sensed data on UV radiation levels, hypotheses concerning the distribution of the skin colors of indigenous peoples relative to UV levels were tested quantitatively in this study for the first time. The major results of this study are: (1) skin reflectance is strongly correlated with absolute latitude and UV radiation levels. The highest correlation between skin reflectance and UV levels was observed at 545 nm, near the absorption maximum for oxyhemoglobin, suggesting that the main role of melanin pigmentation in humans is regulation of the effects of UV radiation on the contents of cutaneous blood vessels located in the dermis. (2) Predicted skin reflectances deviated little from observed values. (3) In all populations for which skin reflectance data were available for males and females, females were found to be lighter skinned than males. (4) The clinal gradation of skin coloration observed among indigenous peoples is correlated with UV radiation levels and represents a compromise solution to the conflicting physiological requirements of photoprotection and vitamin D synthesis. The earliest members of the hominid lineage probably had a mostly unpigmented or lightly pigmented integument covered with dark black hair, similar to that of the modern chimpanzee. The evolution of a naked, darkly pigmented integument occurred early in the evolution of the genus Homo. A dark epidermis protected sweat glands from UV-induced injury, thus insuring the integrity of somatic thermoregulation. Of greater significance to individual reproductive success was that highly melanized skin protected against UV-induced photolysis of folate (Branda & Eaton, 1978, Science201, 625-626; Jablonski, 1992, Proc. Australas. Soc. Hum. Biol.5, 455-462, 1999, Med. Hypotheses52, 581-582), a metabolite essential for normal development of the embryonic neural tube (Bower & Stanley, 1989, The Medical Journal of Australia150, 613-619; Medical Research Council Vitamin Research Group, 1991, The Lancet338, 31-37) and spermatogenesis (Cosentino et al., 1990, Proc. Natn. Acad. Sci. U.S.A.87, 1431-1435; Mathur et al., 1977, Fertility Sterility28, 1356-1360).As hominids migrated outside of the tropics, varying degrees of depigmentation evolved in order to permit UVB-induced synthesis of previtamin D(3). The lighter color of female skin may be required to permit synthesis of the relatively higher amounts of vitamin D(3)necessary during pregnancy and lactation. Skin coloration in humans is adaptive and labile. Skin pigmentation levels have changed more than once in human evolution. Because of this, skin coloration is of no value in determining phylogenetic relationships among modern human groups.",

"query": "how does skin reflectance affect radiation"}

Figure 13: Detailed informations of SCI-DOCS dataset.

An overlord in the English feudal system was a lord of a manor who had subinfeudated a particular manor, estate or fee, to a tenant. The tenant thenceforth owed to the overlord one of a variety of services, usually military service or serjeanty, depending on which form of tenure (i.e. feudal tenancy contract) the estate was "held" under. The highest overlord of all, or paramount lord, was the monarch, who due to his ancestor William the Conqueror's personal conquest of the Kingdom of England, "owned" by inheritance from him all the land in England under allodial title and had no superior overlord, "holding from God and his sword", although certain monarchs, notably King John (1199-1216) purported to grant the Kingdom of England to the Pope, who would thus have become overlord to English monarchs. A paramount lord may thus be seen to occupy the apex of the feudal pyramid, or the root of the feudal tree, and such allodial title is also termed "radical title" (from Latin "radix", root), "ultimate title" and "final title". William the Conqueror immediately set about granting tenancies on his newly won lands, in accordance with feudal principles. The monarch's immediate tenants were the tenants-in-chief, usually military magnates, who held the highest status in feudal society below the monarch. The tenants-in-chief usually held multiple manors or other estates from the monarch, often as feudal barons (or "barons by tenure") who owed their royal overlord an enhanced and onerous form of military service, and subinfeudated most to tenants, generally their own knights or military followers, keeping only a few in demesne. This created a mesne lord - tenant relationship. The knights in turn subinfeudated to their own tenants, creating a further subsidiary mesne lord - tenant relationship. Over the centuries for any single estate the process was in practice repeated numerous times. In early times following the Norman Conquest of 1066 and the establishment of feudalism, land was usually transferred by subinfeudation, rarely by alienation (i.e. sale), which latter in the case of tenants-in-chief required royal licence, and the holder of an estate at any particular time, in order to gain secure tenure, and if challenged by another claimant, needed to prove "devolution of title" evidenced by legal deeds or muniments back up the chain of subinfeudations to a holder whose title was beyond doubt, for example one who had received the estate as a grant by royal charter witnessed and sealed by substantial persons. Although feudal land tenure in England was abolished by the Tenures Abolition Act 1660, in modern English conveyancing law the need to prove devolution of title persisted until recent times, due to a "legal fiction" (grounded in reality) that all land titles were held by the monarch's subjects as a result of a royal grant. Proving devolution of title is no longer necessary since the creation of the land registry and the requirement to compulsorily register all land transactions on this governmental record, which registration provides a virtually unchallengeable and perfectly secure title of ownership.

{"_id": "5a82066d554299676cceb1eb", "text": "What type of system does Rob Donoghue and The Dresden Files Roleplaying Game have in common?"

Figure 14: Detailed informations of HotpotQA dataset.

{"id": "qz_1189--145/145_2602632.txt#0_2", "context": "version of I Will Survive after being sacked from Coronation Street, a strand called The Hopefuls in which viewers came on and did unpleasant things such as snogging an old woman, because, as they said: \"I'd do anything to get on television.\" [PAR] In this post-Big Brother and X Factor world it is hard to imagine now how revolutionary the show was, but in 1990 multichannel TV had hardly started, the independent television production sector was tiny, and the most risque show on TV was Blind Date. People used to go to bed at 11pm, even on Fridays. It was only when Michael Grade , then chief executive of Channel 4, changed the Word's transmission time from 6pm to 11pm a few weeks after it began that the programme became essential post-pub fodder, and it changed viewing habits for ever. [PAR] Social attitudes now reflect the outlook of the generation that watched The Word rather than the ones who complained about it. But in 1990, Britain was a different place, with Margaret Thatcher in her final days as prime minister and only the first flickerings of the optimism and prosperity that would be the abiding spirit of the late-90s and early noughties. Members of the establishment (especially politicians) invariably wore ties. The Word reflected the rift between the laidback attitude of younger people and the establishment, something that no other TV show was then doing. [PAR] Our brief was to appeal to an audience of 16- to 34-year-olds. My goal was to get it talked about the next day – the \"watercooler moment\" as executives would describe it later. Channel 4 , then a different place, supported and encouraged the controversial. Liz Forgan (then Grade's deputy) told us that if she went to a dinner party, and The Word wasn't being attacked ", "answers": {"answer_start": [1242], "text": ["word"]}}

"question": "Which Channel 4 TV show's presenters included Mark Lamarr, Amanda De Cadenet and Terry Christian?"

Figure 15: Detailed informations of TriviaQA dataset.