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processing (Lewis et al., 2021), designed to address several inherent limitations of LLMs - such as hallucination, outdated knowledge, and inadequate domain-specific adaptation (Augenstein et al., 2023) - by incorporating external knowledge.

As LLMs evolve to handle complex long-document tasks, RAG has become essential, driving a shift from fixed-length chunking to more sophisticated strategies. Notable advancements include Meta-Chunking (Zhao et al., 2024) which reformulates segmentation as binary classification, Dense X Retrieval (Chen et al., 2024) which utilizes minimal semantic 'propositions', and Late-Chunking (Günther et al., 2024) which innovatively adopts an 'embed first, then chunk' approach.

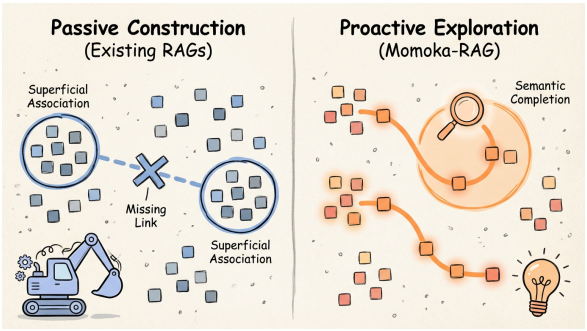

Figure 1: In the figure, the squares denote chunks. The left panel illustrates that existing RAG frameworks, such as RAPTOR and GraphRAG, adhere to fixed rules and rely on statistical data features when organizing knowledge structures. Consequently, they only uncover superficial associations, failing to explore deep links among chunks. In contrast, the right panel demonstrates that Momoka-RAG, driven by semantic completion, successfully identifies deep semantic connections between chunks through active exploration.

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.

To establish the missing connections between chunks, mainstream RAG frameworks have begun attempting to reorganize knowledge structures. Representative works in this direction include the cluster-based RAPTOR and the graph-based GraphRAG. Specifically, RAPTOR adopts a bottom-up approach, employing Gaussian mixture models to cluster chunks and generate summaries to construct a hierarchical tree structure. In contrast, GraphRAG utilizes LLMs to extract entities and relationships within text windows to build a global knowledge graph. Although these

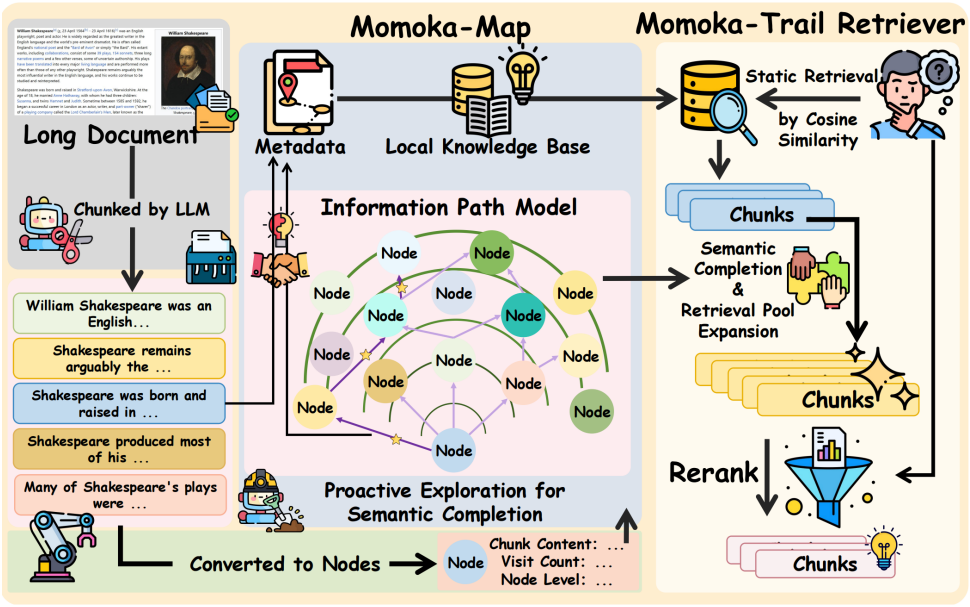

Figure 2: Framework of Momoka-RAG.

methods excel at capturing topic clusters or explicit entity relationships, as illustrated in Figure 1, we argue that they remain trapped in "mechanistic" and "passive" constraints when constructing knowledge structures:

The "mechanistic" is reflected in their reliance on fixed rules to construct structures, lacking adaptive adjustments guided by semantic objectives. For instance, RAPTOR performs clustering based solely on vector similarities, treating all chunks as equally significant while neglecting their distinct roles within the logical chain; similarly, GraphRAG performs undifferentiated entity extraction within fixed windows, lacking active expansion guided by problem-solving goals.

The "passive" stems from their reliance on statistical data features — such as distances in vector space or entity co-occurrence — to generate associations , rather than proactively exploring latent semantic dependencies. Consequently, these frameworks are often limited to capturing superficial associations , struggling to uncover deep cross-paragraph causal relationships or implicit reasoning threads.

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. The specific procedure is shown in Figure 2.

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 design the Momoka-Trail Retriever. This component first obtains initial relevant chunks through static similarity retrieval. Subsequently, it executes a Mixed-Granularity Expansion Strategy, leveraging the path model to construct a semantically rich enhanced candidate pool. Finally, to balance retrieval accuracy and efficiency, we employ a Reranker for fine-grained semantic scoring and ranking, thereby selecting the most precise context based on the user 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 benefit 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 propose the Momoka-Trail Retriever, which fully leverages the information path model to efficiently expand, retrieve, and filter the chunk candidate pool.

(3) We demonstrate competitive performance on carefully selected datasets including Dragonball, SQUAD, HotpotQA, and TriviaQA. This also demonstrates its promising performance in long-document retrieval tasks.

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, to mitigate potential issues of missing subjects or ambiguous references within individual chunks, we first leverage the summarization capability of an LLM to generate a concise global title T_D. Subsequently, utilizing deep linguistic understanding, we dynamically decompose the document into a sequence of semantically coherent chunks and explicitly prepend the global title T_D to the content of each chunk. Please refer to Appendix D for the detailed prompts. 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:

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

Here, $\oplus$ denotes the string concatenation operation. $text_i$ represents the original 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 random selects one such unexplored node c_{new} as a child node for expansion. The reason for adopting this randomized strategy is to increase the breadth of exploration; most importantly,

it simultaneously prevents the formation of cyclic paths.

(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 "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 evaluation function:

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

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 D.

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:

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

(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 $\mathcal{S}$, Top-k K
Output: Retrieved set $\mathcal{F}$

```
// 1. Initial Retrieval with expanded scope
```
$C_{init} \leftarrow \mathcal{R}.\text{retrieve}(Q, 2 \cdot K)$
$C_{aug} \leftarrow \emptyset$
```
// 2. Path-based Expansion Strategies
```
for $c \in C_{init}$ **do**
  $P_c \leftarrow \text{GetPath}(c)$
  `// Expand candidate pool with: Original, Full Path, Nodes, and Pairwise`
  $S_{expand} \leftarrow c \oplus \text{Join}(P_c)$
  for $p \in P_c$ **do**
    $S_{expand} \leftarrow S_{expand} \cup \{p\} \cup \{c \oplus p\}$
  end for
  $C_{aug} \leftarrow C_{aug} \cup S_{expand}$
end for
```
// 3. Global Reranking
```
$\mathcal{F} \leftarrow \mathcal{S}(Q, C_{aug}, K)$
return $\mathcal{F}$

Description: The retrieval process first fetches the top-$2K$ initial chunks via $\mathcal{R}$. For each chunk c, we construct an expanded candidate set C_{aug} by integrating its associated semantic path P_c. This expansion includes the original chunk, the full path context, individual path nodes, and pairwise combinations of the chunk and its nodes. Finally, the Reranker $\mathcal{S}$ scores the entire expanded pool to select the final top-K results $\mathcal{F}$.

3.2 Momoka-Trail Retriever

The Momoka-Map component constructs a high-quality semantic information path model for each chunk within the document. Building on this foundation, the core task of the Momoka-Trail Retriever is to leverage these precomputed paths during the inference phase to identify the most precise context from a broader and semantically enriched candidate pool. Unlike traditional RAG approaches that retrieve only static chunks, we design a "Retrieval -

	Fixed-Length		MSP		PPL		LLM-Chunk		Other Baselines			
Metric	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**LightRAG**	**Late-Chunk**	**RAPTOR**	**D X R**
Dragonball												
Recall	31.62	35.60	85.88	117.04	109.75	167.41	84.14	86.29	31.53	2.23	59.78	3.29
MRR	1.08	1.30	1.83	2.55	2.11	2.68	2.06	2.39	0.65	0.04	1.33	0.11
Precision	59.71	121.60	109.98	246.87	163.76	262.19	121.27	230.80	46.23	3.51	99.18	6.14
F1	0.37	0.55	0.87	1.58	1.26	2.03	0.90	1.25	0.34	0.03	0.68	0.03
HotpotQA												
Recall	262.66	285.45	282.91	284.80	281.02	279.11	277.22	278.48	235.44	11.40	-	237.35
MRR	2.58	2.80	2.79	2.76	2.75	2.69	2.74	2.68	2.11	0.12	-	2.30
Precision	238.61	278.57	256.91	272.82	250.85	265.31	249.54	265.94	184.51	4.05	-	118.78
F1	2.50	2.82	2.69	2.79	2.64	2.72	2.62	2.72	2.06	3.86	-	1.42
TriviaQA												
Recall	189.96	249.03	231.67	254.83	245.56	255.21	244.78	256.36	240.92	198.94	-	160.61
MRR	0.35	2.37	0.44	2.44	0.46	2.38	0.40	2.47	1.98	1.45	-	0.52
Precision	104.25	229.71	134.73	238.41	145.67	234.08	144.68	242.06	163.97	111.20	-	104.60
F1	1.29	2.39	1.64	2.46	1.75	2.44	1.75	2.49	1.91	1.37	-	1.24
SQUAD												
Recall	263.13	270.11	-	-	-	-	277.10	275.14	213.41	70.67	-	191.06
MRR	2.48	2.59	-	-	-	-	2.64	2.61	1.58	0.45	-	1.68
Precision	135.06	253.82	-	-	-	-	144.12	257.41	112.11	34.58	-	96.24
F1	1.64	2.62	-	-	-	-	1.75	2.66	1.38	0.43	-	1.19
NFCORPUS												
Recall	277.64	272.94	281.18	270.98	286.27	271.76	283.93	269.41	206.67	12.95	267.84	271.76
MRR	2.70	2.64	2.72	2.56	2.80	2.58	2.76	2.53	1.89	0.11	2.52	2.63
Precision	250.69	260.39	253.26	255.01	177.21	257.02	255.07	253.88	169.29	4.79	234.64	254.12
F1	2.63	2.66	2.66	2.63	2.66	2.64	2.68	2.61	1.85	0.07	2.50	2.62
SCI-DOCS												
Recall	279.82	271.29	283.85	275.79	285.65	271.52	280.71	269.29	192.83	1.71	287.67	269.06
MRR	2.75	2.65	2.80	2.64	2.82	2.58	2.77	2.55	1.70	0.03	2.79	2.64
Precision	258.43	262.45	264.83	262.14	266.38	257.80	260.69	254.36	144.42	1.02	272.03	254.37
F1	2.68	2.67	2.74	2.69	2.75	2.64	2.70	2.62	1.64	0.01	2.80	2.61

Table 1: **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. In the table, we abbreviate the metrics and framework names, where **PPL**, **MSP**, **D X R**, and **Late-Chunk** stand for **Meta-Chunking-PPL**, **Meta-Chunking-MSP**, **Dense X Retrieval**, and **Late-Chunking**, respectively. Additionally, **Fixed-Length** and **LLM-Chunk** indicate that the chunk sequences input into Momoka-RAG are generated via fixed-length segmentation and LLM-based segmentation, respectively. Furthermore, w/o and w/ denote the methods without and with the processing of Momoka-RAG, respectively.

Expansion - Reranking" pipeline strategy. Specifically, given a user query q, we first employ a vector embedding model to retrieve an initial set of $2 \times k$ relevant chunks, denoted as C_{init}, based on cosine similarity. This ensures sufficient coverage during the initial screening stage. Subsequently, we execute a Mixed-Granularity Expansion Strategy for each retrieved chunk c_i and its associated semantic path P_i. To construct an augmented candidate pool, we not only retain the original chunk to maintain local semantic precision but also concatenate the original chunk with the full path to form a global context. Furthermore, we incorporate each individual node from the path as well as pairwise combinations of "the original chunk and individual nodes." This multi-perspective expansion strategy ensures that both macroscopic narrative clues and microscopic entity associations hidden deep within the path can be effectively captured by the system.

While this expansion strategy significantly enriches the contextual information, it also substantially increases the size of the candidate pool. Faced with such a large-scale candidate set, directly utilizing generative LLMs for processing presents a dilemma: on one hand, adopting item-wise binary filtering leads to unacceptable inference latency, and the overly rigid "Yes/No" decision boundary is prone to inefficiency and the risk of filtering out potentially relevant information; on the other hand, attempting to input all candidates simultaneously into an LLM for listwise ranking not only risks exceeding the model's context window limit but also triggers the "Lost in the Middle" phenomenon due to excessive input length, preventing the model from focusing on key evidence and resulting in a significant degradation in ranking accuracy. Therefore, we employ an efficient Cross-Encoder Reranker to replace the LLM. The reranker computes a continuous, fine-grained semantic relevance score $s(q, d)$ for each candidate document. This "soft" screening mechanism enables more nuanced identification of relevance while maintaining high system respon-

Method	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Relevant	Irrelevant	Wrong
Dragonball								
LightRAG	0.2315	0.1651	0.1101	0.0794	0.0601	48.87	46.74	4.39
PPL	0.3313	0.1425	0.1048	0.0802	0.0630	64.35	33.47	2.05
PPL (+Momoka-RAG)	0.3053	0.1892	0.1436	0.1124	0.0896	81.19	16.28	2.34
TriviaQA								
LightRAG	0.0132	0.0214	0.0191	0.0184	0.0183	-	-	-
PPL	0.0181	0.1018	0.1221	0.1332	0.1393	-	-	-
PPL (+Momoka-RAG)	0.0133	0.0219	0.0197	0.0192	0.0191	-	-	-
SQUAD								
LightRAG	0.1075	0.0576	0.0477	0.0402	0.0355	-	-	-
LLM-Chunk (+Momoka-RAG)	0.1061	0.0630	0.0542	0.0475	0.0429	-	-	-

Table 2: **Comparative Experiments on Generation:** Here, **PPL**, **PPL (+Momoka-RAG)** and **LLM-Chunk (+Momoka-RAG)** represent Meta-Chunking-PPL, Momoka-RAG using Meta-Chunking-PPL as chunking method and Momoka-RAG using LLM as chunking method, respectively.

siveness. Finally, the system ranks the candidate documents based on these scores and selects the top-k documents as the final retrieval result V_{final}, thereby achieving an optimal balance between the accuracy of deep semantic matching and the efficiency of large-scale retrieval. The detailed procedure is presented in Algorithm 1.

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 Chunking, 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 G.

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

$$Metric = Metric@1 + Metric@3 + Metric@5 \quad (6)$$

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

Hyper-parameter. In the comparative experiments, the MCTS iteration count is set to 100, and the maximum rollout length is set to 5. Additionally, we conduct a hyperparameter sensitivity analysis, the results of which are presented in Table 4. Throughout all experiments, α is set to 3, β is set to 2, and γ and δ are set to 1.

Experiment Organization. We designed the following experiments: (1) Comparative experiments on retrieval and generation performance; (2) Ablation studies; (3) Sensitivity analysis of critical hyper-parameters; and (4) A specific investigation into the MCTS reward mechanism design. Due to space limitations, items (3) and (4) are presented in the Appendix E and F.

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 1. Due to space limitations, complete experimental results can be found in Appendix H.

As Momoka-RAG functions as a plug-and-play framework capable of processing arbitrary sequences of pre-segmented chunks, our experiments in this section incorporate four segmentation methods: Fixed-Length, LLM-based, Meta-Chunking-MSP, and Meta-Chunking-PPL. It is worth noting that the Fixed-Length and LLM-based methods operate at the paragraph level. In contrast, the Meta-Chunking methods (MSP and PPL) process the entire document globally. Since the latter approach disrupts the original paragraph structure, rendering paragraph indices invalid, we treat paragraph IDs as equivalent to sentence IDs when processing the results from Meta-Chunking.

Due to space limitations, the detailed result analysis is presented in Appendix A.

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**). Table 2 presents the experimental results.

Due to space limitations, the detailed result analysis is presented in Appendix B.

For details regarding the corresponding prompts and metric descriptions, please refer to Appendix I.

4.3 Ablation Studies

Momoka-RAG is constructed by integrating the **Momoka-Trail Retriever** component with the **Momoka-Map** chunking method. To validate the effectiveness of each component within the framework, we conducted ablation studies on the corresponding datasets, as shown in Table 3.

Due to space limitations, the detailed result analysis is presented in Appendix C.

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 semantic 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) 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.

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 Analysis of Comparative Experiments on Retrieval Performance

Although Momoka-RAG significantly enhances retrieval performance across most scenarios, we observed that under specific combinations of segmentation methods and datasets, the Recall metric exhibits fluctuations, whereas Precision demonstrates a substantial improvement. We provide an in-depth analysis of this phenomenon from three dimensions: the coherence of segmentation granularity, the effectiveness of positional encoding, and the precision trade-off.

Complementary Effects of Segmentation Granularity and Contextual Coherence. The magnitude of performance gains achieved by Momoka-RAG across different chunking methods is closely correlated with the contextual coherence of the original chunks:

Local Segmentation (Fixed-Length, LLM-Chunk): These methods typically treat the "paragraph" as the processing unit, restricting chunks strictly within paragraph boundaries. This fragmentation results in severe semantic gaps. Momoka-RAG utilizes MCTS to proactively construct cross-paragraph semantic paths, effectively acting as a "semantic adhesive." Consequently, it achieves the most significant performance leaps on such chunking methods.

Global Segmentation (Meta-Chunking MSP/PPL): These approaches input the entire document into an LLM for processing, generating chunks that are inherently cross-paragraph, possessing high native coherence and larger granularity. Since the original chunks already encapsulate sufficient context, the marginal utility of Momoka-RAG's "semantic completion" naturally diminishes. In certain information-dense documents, excessive path expansion may even dilute the core semantics of the original chunks.

Effectiveness Analysis of Structural Prior Information. Momoka-RAG's reward function relies on the spatial positional information of chunks (paragraph number p_i and sentence number s_i) to

assist in judging semantic distance. For Fixed-Length and LLM-Chunk methods, where the paragraph structure is distinct, Momoka-Map can precisely leverage these positional priors. However, for Meta-Chunking, the disruption of the original paragraph structure results in the loss of paragraph numbering information. In our experiments, we treated paragraph numbers as equivalent to sentence numbers for this case, which effectively weakened the constraint of the Structural Prior. In structurally rigorous scientific literature such as SCI-DOCS, this blurring of structural information may lead to deviations in the MCTS exploration direction, constituting one of the technical reasons for the insignificant improvement or even decline in Recall under this combination.

"High-Precision" Oriented Filtering Mechanism. It is worth noting that on datasets like NFCORPUS and SCI-DOCS, despite a slight decline in Recall when combined with Meta-Chunking, Precision often achieved substantial improvements. This indicates that Momoka-RAG essentially functions as a high-intensity semantic filter. In scientific and medical retrieval, traditional vector retrieval often recalls a large volume of "pseudo-relevant" documents that contain matching keywords but are semantically unrelated (leading to high Recall but low Precision). Through the construction of semantic paths, Momoka-RAG greatly purifies the candidate set—although the noise introduced by path expansion may cause some marginally relevant documents to be missed—ensuring that the retained documents are highly semantically matched.

Adaptability Differences Between Narrative and Factual Texts. The characteristics of the datasets themselves also amplify the aforementioned phenomena. In strongly narrative texts like Dragonball, cross-paragraph semantic dependency is the core pain point; hence, Momoka-RAG demonstrates optimal effectiveness. Conversely, in NFCORPUS and SCI-DOCS, the texts consist mostly of high-density independent factual statements replete with technical terminology. In such scenarios, the "knowledge atoms" segmented by Meta-Chunking are already sufficiently complete. When superimposed with Momoka-RAG's path expansion, although precision is enhanced, the introduction of additional context inevitably incurs a certain degree of recall loss.

B Detailed Analysis of Comparative Experiments on Generation Quality

Experimental results demonstrate that the generative advantages of Momoka-RAG are primarily concentrated in complex contextual tasks. In long-range narrative scenarios such as Dragonball, benefiting from the complete evidence chains constructed by MCTS, the responses generated by the model significantly outperform baseline methods in terms of logical relevance and information completeness, effectively addressing the issue of "contextual fragmentation" often encountered in traditional retrieval. Conversely, on tasks emphasizing short fact extraction, such as SQUAD and TriviaQA, Momoka-RAG exhibits mixed results or slight fluctuations compared to baselines. This is mainly because the framework tends to generate more detailed and explanatory verbose responses based on rich context, rather than simple keyword extraction. Consequently, it does not hold an advantage on N-gram metrics that favor literal phrase-level matching, although this does not imply a deficiency in actual semantic accuracy.

C Detailed Analysis of Ablation Studies

A comparison of the experimental results between Base and MM reveals that this mechanism yields significant performance improvements across the vast majority of datasets. In narrative or QA-oriented datasets such as Dragonball and TriviaQA, the MM configuration demonstrates substantial growth in both Precision and Recall compared to Base. This is attributed to the fact that original chunks are often replete with ambiguous references such as "he" or "that place." By forcibly injecting the global document title T_D, Momoka-Map assigns an explicit semantic label to each chunk, enabling the retrieval model to precisely localize contextually relevant entities, thereby significantly reducing the false positive rate. The improvement brought by MM is particularly dramatic for mechanical segmentation methods like Fixed-Length. This demonstrates that simple Metadata Augmentation can effectively mitigate the loss of context caused by forced truncation, allowing isolated paragraphs to regain their positioning capability within the document. However, in highly structured scientific literature such as SCI-DOCS, we observe a slight decline in Recall from Base to MM. This may be because scientific titles often consist of generalized domain vocabulary, which offers limited

assistance in distinguishing specific experimental details within the document. Conversely, this may interfere with fine-grained retrieval discriminability by increasing the similarity overlap of embeddings.

By comparing the experimental results of MM and Momoka, we can further analyze the practical effectiveness of the Momoka-Trail Retriever in utilizing MCTS paths to expand the candidate pool. In the TriviaQA and Dragonball datasets, the Momoka configuration achieves a further performance leap over MM. This validates the effectiveness of MCTS paths — representing not merely simple relevance matching, but a logical process of "following the clues." The Retriever pulls clues scattered across different locations in the document into the candidate pool simultaneously via these paths, greatly enhancing the completeness of evidence for complex queries. Conversely, in datasets with extremely high factual density, such as NFCORPUS and SCI-DOCS, path expansion introduces complex trade-offs. Under certain configurations, while Recall remains relatively stable, Precision shows a significant improvement. This suggests that path expansion imposes stricter semantic constraints; the Retriever leverages path information to filter out pseudo-relevant chunks that are "seemingly related but logically incoherent," effectively functioning as a high-precision filter. However, in combinations like NFCORPUS-PPL, we observe a decline in both Recall and Precision from MM to Momoka. This occurs because, in documents where knowledge points are highly independent, forcibly establish MCTS paths may connect unrelated concepts, thereby introducing a large amount of irrelevant information during retrieval that crowds out the Top-K space. This indicates that for scenarios with extremely high densities of fragmented knowledge, excessive path expansion may yield diminishing marginal utility or even negative returns. In such cases, retaining only the context augmentation provided by Momoka-Map may be the optimal choice.

D Detailed Prompt in Momoka-RAG

In Figures 3, 4 and 5, we have presented the prompt details of the Momoka-RAG.

E Sensitivity Analysis of Critical Hyper-parameters

During the construction of the Momoka-Map, two hyper-parameters of MCTS — the Maximum Iteration Count (N_{iter}) and the Maximum Rollout Steps ($L_{rollout}$) — directly govern the breadth and depth of semantic path exploration. To investigate the impact of these two parameters on retrieval performance, we conduct comparative experiments using seven distinct parameter combinations in Table 4. Given that Recall and Precision metrics exhibit consistent trends across the vast majority of comparisons, we focus exclusively on the analysis of the Recall metric in this section.

Analyzing the impact of Rollout steps with fixed iteration counts: In the high-iteration regime ($N_{iter} = 100$), the Recall at $L_{rollout} = 1$ significantly outperforms that at $L_{rollout} = 3$. This implies that when search breadth is sufficient, deep rollouts inadvertently act as a source of noise. Due to the decaying nature of semantic associations, as simulation steps increase, chunks at the end of the path may experience Semantic Drift, deviating from the original context of the root node. Consequently, the immediate reward signal derived from evaluating only the direct neighbor ($L_{rollout} = 1$) remains the most pristine and effective. In the medium-iteration regime ($N_{iter} = 50$), the Recall at $L_{rollout} = 5$ is slightly higher than at $L_{rollout} = 1$ (and $L_{rollout} = 3$). When the search breadth is insufficient to cover all critical branches, increasing depth exploration may serendipitously uncover hidden semantic associations, thereby compensating for the lack of breadth to a certain extent. However, this proves to be an inefficient strategy, yielding marginal gains compared to the dividends obtained from increasing the iteration count. Conversely, in the low-iteration regime ($N_{iter} = 20$), $L_{rollout} = 3$ outperforms $L_{rollout} = 5$. This indicates that under severely constrained computational budgets, excessively deep rollouts may deplete the resources available for individual simulations, and the resulting high variance leads to highly unstable return estimates.

Analyzing the impact of iteration counts with fixed Rollout steps: In shallow simulations ($L_{rollout} = 1$), increasing the iteration count from $N_{iter} = 50$ to $N_{iter} = 100$ achieves a qualitative leap in Recall. This constitutes the most decisive finding of the comparison, demonstrating that in semantic path construction tasks, breadth-first is the unequivocally dominant strategy. Only with a sufficient number of samples can the UCT algorithm of MCTS effectively balance exploration and exploitation, thereby covering non-obvious yet high-quality semantic paths. In deep simulations ($L_{rollout} = 3$),

	Fixed-Length			Meta-Chunking-MSP			Meta-Chunking-PPL			LLM-Chunk		
Metric	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**
Dragonball												
Recall	31.62	36.14	35.60	85.88	100.42	117.04	109.75	137.27	167.41	84.14	67.88	86.29
MRR	1.08	1.24	1.30	1.83	2.25	2.55	2.11	2.32	2.68	2.06	1.90	2.39
Precision	59.71	91.89	121.60	109.98	184.10	246.87	163.76	143.24	262.19	121.27	116.03	230.80
F1	0.37	0.50	0.55	0.87	1.25	1.58	1.26	1.11	2.03	0.90	0.77	1.25
NFCORPUS												
Recall	277.64	282.35	272.94	281.18	283.93	270.98	286.27	287.09	271.76	283.93	281.96	269.41
MRR	2.70	2.78	2.64	2.72	2.80	2.56	2.80	2.84	2.58	2.76	2.79	2.53
Precision	250.69	270.76	260.39	253.26	272.06	255.01	177.21	271.20	257.02	255.07	271.48	253.88
F1	2.63	2.76	2.66	2.66	2.78	2.63	2.66	2.79	2.64	2.68	2.77	2.61
HotpotQA												
Recall	262.66	285.44	285.45	282.91	286.08	284.80	281.02	291.14	279.11	277.22	284.81	278.48
MRR	2.58	2.84	2.80	2.79	2.84	2.76	2.75	2.87	2.69	2.74	2.82	2.68
Precision	238.61	274.52	278.57	256.91	274.51	272.82	250.85	186.32	265.31	249.54	271.31	265.94
F1	2.50	2.80	2.82	2.69	2.80	2.79	2.64	2.81	2.72	2.62	2.78	2.72
SCI-DOCS												
Recall	279.82	274.43	271.29	283.85	282.07	275.79	285.65	282.06	271.52	280.71	282.73	269.29
MRR	2.75	2.73	2.65	2.80	2.80	2.64	2.82	2.81	2.58	2.77	2.81	2.55
Precision	258.43	265.45	262.45	264.83	271.93	262.14	266.38	272.44	257.80	260.69	271.70	254.36
F1	2.68	2.70	2.67	2.74	2.77	2.69	2.75	2.77	2.64	2.70	2.77	2.62
SQUAD												
Recall	263.13	271.51	270.11	-	-	-	-	-	-	277.10	281.00	275.14
MRR	2.48	2.60	2.59	-	-	-	-	-	-	2.64	2.71	2.61
Precision	135.06	142.17	253.82	-	-	-	-	-	-	144.12	148.90	257.41
F1	1.64	1.72	2.62	-	-	-	-	-	-	1.75	1.80	2.66
TriviaQA												
Recall	189.96	226.25	249.03	231.67	243.63	254.83	245.56	257.14	255.21	244.78	249.04	256.36
MRR	0.35	0.34	2.37	0.44	0.43	2.44	0.46	0.45	2.38	0.40	0.44	2.47
Precision	104.25	159.56	229.71	134.73	167.62	238.41	145.67	169.65	234.08	144.68	166.66	242.06
F1	1.29	1.84	2.39	1.64	1.95	2.46	1.75	1.99	2.44	1.75	1.95	2.49

Table 3: **Ablation Studies:** Due to the presence of sensitive or unsafe content in the original documents of the datasets, LLMs cannot be utilized for chunking in the Meta-Chunking method. In the table, framework names are abbreviated, where **Base**, **MM**, and **Momoka** represent using the chunking method solely, processing the pre-chunked sequence through Momoka-Map, and processing the pre-chunked sequence through Momoka-RAG, respectively.

performance exhibits non-monotonic variations as iterations increase. For instance, the performance dip at $N_{iter} = 50$ suggests a potential incompatibility zone between medium-scale search and deep simulation. At this stage, the search tree is neither sufficiently broad (prone to local optima) nor are the simulation signals sufficiently accurate, causing the model to falter. Only when the iteration count is further increased to 100 does the robustness of the search tree once again overcome the noise introduced by deep simulations.

Due to constraints on computational resources and experimental timelines, we did not explore larger settings for iteration counts or rollout depths. Current experimental results indicate that the model achieves superior performance at $Iteration = 100$, while exhibiting a distinct trend of diminishing marginal returns. This suggests that further increasing computational overhead would not yield significant improvements in metrics.

In summary, the expansion of search breadth and the resulting performance gains driven by increased iteration counts precisely reflect the core essence of 'Proactive Exploration' in Momoka-RAG. This constitutes the fundamental innovation that enables our framework to reconstruct knowledge structures and overcome the bottlenecks of long-document retrieval.

F Analysis of MCTS Reward

The reward function r serves as a compass guiding MCTS to perform effective searches within the semantic space. In this study, we compare two different reward mechanism designs: the original mechanism (denoted as w/o), which focuses exclusively on the initial starting point, and the modified mechanism (denoted as w/), which accounts for the complete evolutionary path. Experimental results are presented in Table 5. Both mechanisms maintain consistent physical constraints by calculating

Metric	Iter 20 / Roll 3	Iter 20 / Roll 5	Iter 50 / Roll 1	Iter 50 / Roll 3	Iter 50 / Roll 5	Iter 100 / Roll 1	Iter 100 / Roll 3
Recall	80.25	79.13	78.48	78.45	79.76	88.71	86.29
MRR	2.35	2.31	2.33	2.33	2.34	2.42	2.39
Precision	223.27	220.60	222.95	221.76	224.01	235.13	230.80
F1	1.17	1.15	1.16	1.15	1.17	1.28	1.25

Table 4: **Sensitivity Analysis of Critical Hyper-parameters:** We conduct a hyper-parameter sensitivity analysis on the Dragonball dataset. In the table, the notation **Iter A/Roll B** denotes that the number of iterations is set to A and the maximum rollout steps are set to B.

Dataset	Method	Recall	Precision	MRR	F1
Dragonball	w/	86.56	234.91	2.43	1.26
	w/o	86.29	230.80	2.39	1.25
NFCORPUS	w/	270.19	254.46	2.53	2.62
	w/o	269.41	253.88	2.53	2.61
SCI-DOCS	w/	270.63	255.18	2.56	2.63
	w/o	269.29	254.36	2.55	2.62
TriviaQA	w/	259.46	242.59	2.49	2.51
	w/o	256.36	242.06	2.47	2.49

Table 5: **Analysis of MCTS Reward:** In the table, all experimental parameter settings are consistent with those in the main text. The notations **w/** and **w/o** denote whether the reward mechanism within the MCTS process is modified. Due to the presence of sensitive content in certain datasets and recent updates to the API, the method was unable to function properly. Consequently, the experimental results for these datasets are not presented.

the physical distance of the new node relative to the initial chunk, thereby preventing the search scope from diverging excessively within the document. However, they exhibit a fundamental difference in the logic used to determine semantic consistency: in the original mechanism, the LLM merely evaluates whether the new chunk is semantically relevant to the initial chunk; whereas in the modified mechanism, the LLM assesses whether the new chunk constitutes a reasonable logical continuation of the current complete path context. The detailed prompts can be found in Appendix D. The formula for the modified reward mechanism is defined as follows:

$$
\begin{aligned}
r = \mathbb{I}_{LLM}(\text{concat}(\mathcal{P}_{context}), c_{new}) \\
+ \alpha \cdot \frac{1}{|p_i - p_{\text{new}}| + \gamma} \\
+ \beta \cdot \frac{1}{|s_i - s_{\text{new}}| + \delta},
\end{aligned}
\tag{7}
$$

Unlike the original mechanism, here we concatenate the textual content of the complete path from the root node to the current node, denoted as $\mathcal{P}_{context}$, to serve as the input context for the LLM. The function $\mathbb{I}_{LLM}$ indicates whether the new chunk c_{new} constitutes a reasonable logical continuation of this complete logical flow (returning 1 if valid, and 0 otherwise).

Based on the experimental results, we focus our analysis exclusively on the Dragonball and TriviaQA datasets, where significant comparative improvements are observed.

In the Dragonball corpus, technical terminology frequently recurs across various perspectives of financial analysis. This high degree of terminological overlap renders baseline methods (**w/o**), which rely solely on local relevance, highly susceptible to confusing logical attributions across different timelines or business segments. The modified reward mechanism (**w/**), by enforcing the inclusion of full path context ($\mathcal{P}_{context}$), imposes a strict logical consistency constraint on MCTS. It shifts the focus from mere lexical matching between nodes to scrutinizing whether a new node seamlessly continues the financial analysis logic of the preceding path. This mechanism effectively filters out "pseudo-relevant" information—chunks that share identical terms but are misplaced within the logical flow—thereby enhancing retrieval precision while maintaining stable recall. This finding is consistent with the high sensitivity to long-range logical dependencies observed for this dataset in our main experiments.

Conversely, in datasets like TriviaQA which target open-domain question answering, the characteristics favor discrete "short fact extraction." Questions often necessitate gathering evidence across multiple non-contiguous paragraphs, where connections between knowledge points are relatively loose. In this scenario, the modified reward mechanism exhibits a distinctly different operational mode: it functions as a "bridge" for knowledge association rather than a "filter." By evaluating the overall rationality of the path, the model identifies distal nodes that, despite being semantically

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

Table 6: Average chunk length of each dataset.

distant from the starting node in a literal sense, constitute reasonable extensions within the knowledge inference chain. This tolerance for large semantic jumps enables MCTS to construct evidence paths with broader coverage, successfully reaching deeply hidden knowledge fragments. This explains why the modified mechanism primarily yields a significant surge in Recall on TriviaQA, demonstrating that full path context effectively mitigates premature search truncation caused by excessive semantic distances in divergent knowledge tasks.

G 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 6, 7, 8, 9, 10 and 11. The average chunk length for each dataset is shown in Table 6.

H Detailed Information of Comparative Experiments

This section presents the comprehensive experimental results of the comparative experiments. Across all tables, all abbreviations remain consistent with those used in the main text. Please refer to Tables 7, 8, 9, 10, 11 and 12 for details.

I 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 covers 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].

J Detailed Information of Ablation Studies

This section presents the comprehensive experimental results of the ablation studies. Across all tables, all abbreviations remain consistent with those used in the main text. Please refer to Tables 13, 14, 15, 16, 17 and 18 for details.

K Detailed Information of Sensitivity Analysis of Critical Hyper-parameters

This section presents the comprehensive experimental results of the sensitivity analysis of critical hyper-parameters. Across all tables, all abbreviations remain consistent with those used in the main text. Please refer to Table 19 for details.

L Detailed Information of Analysis of MCTS Reward

This section presents the comprehensive experimental results of the analysis of mcts reward. Across all tables, all abbreviations remain consistent with those used in the main text. Please refer to Tables 20, 21, 22 and 23 for details.

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

```
You are a master of text segmentation. You are now provided with a natural
paragraph of text that requires segmentation. Please merge and split at the
sentence level while respecting the original logic, semantic connections, and
coherence. Note: Do not alter any of the original content; your sole task is
segmentation.

        **Text Paragraph**: {item}

        Notes:
        1. Output the result as a list in the format ["Chunk1", "Chunk2", ...].
The number of chunks is up to you. Output ONLY the list and nothing else.
        2. Do not split originally complete sentences.
        3. Segmentation is not simply isolating individual sentences; you must
identify semantically coherent combinations of sentences to form a Chunk. The
results must respect the logic and semantic links of the original text.

        **Output Format**: ["Chunk1", "Chunk2", ...]
```

Figure 3: Detailed prompt in Momoka-Map for llm chunk.

```
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 4: Detailed prompt in Momoka-Map.

```
        You are a master of information judgment. You are provided with a
**Preceding Text** (concatenated from multiple chunks) and a **New Chunk**.
        Please determine if the **New Chunk** is a reasonable semantic
continuation or completion of the **Preceding Text** (e.g., logical coherence,
subject continuity, causal relationship, or thematic elaboration).

        If the New Chunk follows the Preceding Text well, return 1; if there is
a logical break or it is irrelevant, return 0. Do not output anything else.

        **Preceding Text**:
        {context_text}

        **New Chunk**:
        {new_chunk.text}

        Output Format: 1/0
```

Figure 5: Detailed prompt in Momoka-Map with reward change.

	Fixed-Length		MSP		PPL		LLM-Chunk		Other Baselines			
Metric	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**LightRAG**	**Late-Chunk**	**RAPTOR**	**D X R**
Top-1												
Recall	7.53	9.86	18.35	29.93	26.87	43.32	14.87	22.66	6.61	0.52	14.40	0.75
MRR	0.33	0.41	0.56	0.81	0.68	0.87	0.54	0.76	0.24	0.02	0.53	0.03
Precision	32.55	41.20	55.95	80.57	68.06	86.78	54.32	76.09	23.70	1.83	53.31	3.26
F1	0.12	0.16	0.28	0.44	0.39	0.58	0.23	0.35	0.10	0.01	0.23	0.01
Top-3												
Recall	11.01	11.86	30.12	39.98	37.84	58.14	31.51	28.85	10.92	0.75	20.89	1.18
MRR	0.37	0.44	0.63	0.87	0.71	0.91	0.75	0.80	0.22	0.01	0.42	0.04
Precision	15.87	40.66	31.06	83.35	54.26	88.00	38.93	77.35	12.72	0.95	26.50	1.70
F1	0.13	0.18	0.31	0.54	0.45	0.70	0.35	0.42	0.12	0.01	0.23	0.01
Top-5												
Recall	13.08	13.88	37.41	47.13	45.04	65.95	37.76	34.78	14.00	0.96	24.49	1.36
MRR	0.38	0.45	0.64	0.87	0.72	0.90	0.76	0.83	0.19	0.01	0.37	0.04
Precision	11.29	39.74	22.97	82.95	41.44	87.41	28.02	77.36	9.81	0.73	19.37	1.18
F1	0.12	0.21	0.28	0.60	0.43	0.75	0.32	0.48	0.12	0.01	0.22	0.01

Table 7: Performance Comparison on Dragonball Dataset (Top-1, Top-3, Top-5).

	Fixed-Length		MSP		PPL		LLM-Chunk		Other Baselines			
Metric	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**LightRAG**	**Late-Chunk**	**RAPTOR**	**D X R**
Top-1												
Recall	87.84	84.31	88.24	80.00	91.76	81.57	90.20	79.61	58.04	2.75	80.39	85.49
MRR	0.88	0.84	0.88	0.80	0.92	0.82	0.90	0.80	0.58	0.03	0.80	0.85
Precision	87.84	84.31	88.24	80.00	91.76	81.57	90.20	79.61	58.04	2.75	80.39	85.49
F1	0.88	0.84	0.88	0.80	0.92	0.82	0.90	0.80	0.58	0.03	0.80	0.85
Top-3												
Recall	94.90	94.12	95.69	94.51	95.69	95.29	96.08	94.51	72.55	5.10	91.76	92.94
MRR	0.91	0.90	0.91	0.88	0.93	0.89	0.93	0.86	0.65	0.04	0.85	0.89
Precision	83.40	87.84	84.71	87.32	84.71	87.06	84.71	86.27	56.21	1.02	77.27	85.49
F1	0.89	0.91	0.90	0.91	0.90	0.91	0.90	0.90	0.63	0.02	0.84	0.89
Top-5												
Recall	94.90	94.51	97.25	96.47	98.82	94.90	97.65	95.29	76.08	5.10	95.69	93.33
MRR	0.91	0.90	0.92	0.88	0.94	0.88	0.93	0.87	0.66	0.04	0.87	0.88
Precision	79.45	88.24	80.31	87.69	0.74	88.39	80.16	88.00	55.04	1.02	76.98	83.14
F1	0.86	0.91	0.88	0.92	0.85	0.92	0.88	0.92	0.64	0.02	0.85	0.88

Table 8: Performance Comparison on NFCORPUS Dataset (Top-1, Top-3, Top-5).

	Fixed-Length		MSP		PPL		LLM-Chunk		Other Baselines			
Metric	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**LightRAG**	**Late-Chunk**	**RAPTOR**	**D X R**
Top-1												
Recall	85.44	92.41	91.77	87.97	90.51	83.54	90.51	85.44	63.92	3.80	-	75.32
MRR	0.85	0.92	0.92	0.88	0.91	0.84	0.91	0.85	0.64	0.04	-	0.75
Precision	85.44	92.41	91.77	87.97	90.51	83.54	90.51	85.44	63.92	0.04	-	75.32
F1	0.85	0.92	0.92	0.88	0.91	0.84	0.91	0.85	0.64	3.80	-	0.75
Top-3												
Recall	86.71	96.84	95.57	98.10	94.94	97.47	94.30	96.20	84.18	3.80	-	80.38
MRR	0.85	0.95	0.94	0.94	0.92	0.92	0.92	0.91	0.73	0.04	-	0.77
Precision	77.22	93.25	85.65	92.19	85.02	89.87	83.97	89.87	61.35	2.11	-	27.00
F1	0.82	0.95	0.90	0.95	0.90	0.94	0.89	0.93	0.71	0.03	-	0.40
Top-5												
Recall	90.51	96.20	95.57	98.73	95.57	98.10	92.41	96.84	87.34	3.80	-	81.65
MRR	0.88	0.93	0.93	0.94	0.93	0.93	0.91	0.91	0.74	0.04	-	0.78
Precision	75.95	92.91	79.49	92.66	75.32	91.90	75.06	90.63	59.24	1.90	-	16.46
F1	0.83	0.95	0.87	0.96	0.84	0.95	0.83	0.94	0.71	0.03	-	0.27

Table 9: Performance Comparison on HotpotQA Dataset (Top-1, Top-3, Top-5).

	Fixed-Length		MSP		PPL		LLM-Chunk		Other Baselines			
Metric	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**LightRAG**	**Late-Chunk**	**RAPTOR**	**D X R**
Top-1												
Recall	90.36	88.79	92.15	86.55	93.50	81.61	91.70	82.74	50.90	0.37	93.05	86.77
MRR	0.90	0.89	0.92	0.87	0.94	0.82	0.92	0.83	0.51	0.01	0.93	0.87
Precision	90.36	88.79	92.15	86.55	93.50	81.61	91.70	82.74	50.90	0.67	93.05	86.77
F1	0.90	0.89	0.92	0.87	0.94	0.82	0.92	0.83	0.51	0.01	0.93	0.87
Top-3												
Recall	94.84	90.80	95.29	94.62	95.96	95.07	94.84	92.38	68.16	0.67	97.09	90.81
MRR	0.92	0.88	0.94	0.89	0.95	0.89	0.93	0.86	0.59	0.01	0.92	0.89
Precision	85.87	86.62	87.97	87.74	88.71	87.89	87.29	84.53	47.52	0.22	90.43	84.01
F1	0.90	0.89	0.91	0.91	0.92	0.91	0.91	0.88	0.56	0.00	0.94	0.87
Top-5												
Recall	94.62	91.70	96.41	94.62	96.19	94.84	94.17	94.17	73.77	0.67	97.53	91.48
MRR	0.92	0.88	0.94	0.89	0.94	0.88	0.92	0.86	0.60	0.01	0.94	0.89
Precision	82.20	87.04	84.71	87.85	84.17	88.30	81.70	87.09	46.00	0.13	88.55	83.59
F1	0.88	0.89	0.90	0.91	0.90	0.91	0.88	0.90	0.57	0.00	0.93	0.87

Table 10: Performance Comparison on SCI-DOCS Dataset (Top-1, Top-3, Top-5).

	Fixed-Length		MSP		PPL		LLM-Chunk		Other Baselines			
Metric	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**LightRAG**	**Late-Chunk**	**RAPTOR**	**D X R**
Top-1												
Recall	79.05	86.31	-	-	-	-	84.64	87.15	56.27	16.48	-	50.84
MRR	0.79	0.86	-	-	-	-	0.85	0.87	0.56	0.16	-	0.51
Precision	79.05	86.31	-	-	-	-	84.64	87.15	56.27	16.48	-	50.84
F1	0.79	0.86	-	-	-	-	0.85	0.87	0.56	0.16	-	0.51
Top-3												
Recall	90.78	90.78	-	-	-	-	94.97	93.02	75.51	25.70	-	66.76
MRR	0.84	0.86	-	-	-	-	0.89	0.88	0.53	0.16	-	0.58
Precision	33.89	83.71	-	-	-	-	35.85	85.85	30.82	10.61	-	26.35
F1	0.49	0.87	-	-	-	-	0.52	0.89	0.44	0.15	-	0.38
Top-5												
Recall	93.30	93.02	-	-	-	-	97.49	94.97	81.63	28.49	-	73.46
MRR	0.85	0.86	-	-	-	-	0.90	0.86	0.49	0.13	-	0.59
Precision	22.12	83.80	-	-	-	-	23.63	84.41	25.02	7.49	-	19.05
F1	0.36	0.88	-	-	-	-	0.38	0.89	0.38	0.12	-	0.30

Table 11: Performance Comparison on SQUAD Dataset (Top-1, Top-3, Top-5).

	Fixed-Length		MSP		PPL		LLM-Chunk		Other Baselines			
Metric	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**w/o**	**w/**	**LightRAG**	**Late-Chunk**	**RAPTOR**	**D X R**
Top-1												
Recall	45.56	75.29	59.85	79.15	68.34	78.38	64.86	81.85	67.57	46.72	-	37.83
MRR	0.10	0.75	0.13	0.79	0.14	0.78	0.11	0.82	0.68	0.47	-	0.38
Precision	45.56	75.29	59.85	79.15	68.34	78.38	64.86	81.85	67.57	46.72	-	37.83
F1	0.46	0.75	0.60	0.79	0.68	0.78	0.65	0.82	0.68	0.47	-	0.38
Top-3												
Recall	67.18	86.10	82.24	87.64	86.10	88.80	87.26	86.87	85.71	71.43	-	56.76
MRR	0.12	0.81	0.15	0.83	0.16	0.81	0.14	0.83	0.65	0.51	-	0.07
Precision	32.05	77.35	41.44	80.57	45.05	79.41	46.46	81.21	51.56	34.36	-	34.88
F1	0.43	0.81	0.55	0.84	0.59	0.84	0.61	0.84	0.64	0.46	-	0.43
Top-5												
Recall	77.22	87.64	89.58	88.04	91.12	88.03	92.66	87.64	87.64	80.79	-	66.02
MRR	0.13	0.81	0.16	0.82	0.17	0.78	0.15	0.82	0.65	0.47	-	0.07
Precision	26.64	77.07	33.44	78.69	32.28	76.29	33.36	79.00	44.84	30.12	-	31.89
F1	0.40	0.82	0.49	0.83	0.48	0.82	0.49	0.83	0.59	0.44	-	0.43

Table 12: Performance Comparison on TriviaQA Dataset (Top-1, Top-3, Top-5).

	Fixed-Length			Meta-Chunking-MSP			Meta-Chunking-PPL			LLM-Chunk		
Metric	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**
Top-1												
Recall	7.53	8.80	9.86	18.35	22.00	29.93	26.87	28.80	43.32	14.87	18.87	22.66
MRR	0.33	0.37	0.41	0.56	0.68	0.81	0.68	0.73	0.87	0.54	0.70	0.76
Precision	32.55	37.44	41.20	55.95	68.26	80.57	68.06	72.74	86.78	54.32	69.58	76.09
F1	0.12	0.14	0.16	0.28	0.33	0.44	0.39	0.41	0.58	0.23	0.30	0.35
Top-3												
Recall	11.01	12.96	11.86	30.12	36.66	39.98	37.84	48.99	58.14	31.51	22.49	28.85
MRR	0.37	0.43	0.44	0.63	0.75	0.87	0.71	0.79	0.91	0.75	0.59	0.80
Precision	15.87	18.68	40.66	31.06	38.08	83.35	54.26	40.90	88.00	38.93	27.20	77.35
F1	0.13	0.15	0.18	0.31	0.37	0.54	0.45	0.45	0.70	0.35	0.25	0.42
Top-5												
Recall	13.08	14.38	13.88	37.41	41.76	47.13	45.04	59.48	65.95	37.76	26.52	34.78
MRR	0.38	0.44	0.45	0.64	0.82	0.87	0.72	0.80	0.90	0.76	0.61	0.83
Precision	11.29	35.77	39.74	22.97	77.76	82.95	41.44	29.60	87.41	28.02	19.25	77.36
F1	0.12	0.21	0.21	0.28	0.54	0.60	0.43	0.25	0.75	0.32	0.22	0.48

Table 13: Ablation Studies on the Dragonball dataset for Top-1, Top-3, and Top-5 retrieval.

	Fixed-Length			Meta-Chunking-MSP			Meta-Chunking-PPL			LLM-Chunk		
Metric	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**
Top-1												
Recall	87.84	89.80	84.31	88.24	92.16	80.00	91.76	93.73	81.57	90.20	92.16	79.61
MRR	0.88	0.90	0.84	0.88	0.92	0.80	0.92	0.94	0.82	0.90	0.92	0.80
Precision	87.84	89.80	84.31	88.24	92.16	80.00	91.76	93.73	81.57	90.20	92.16	79.61
F1	0.88	0.90	0.84	0.88	0.92	0.80	0.92	0.94	0.82	0.90	0.92	0.80
Top-3												
Recall	94.90	96.86	94.12	95.69	95.69	94.51	95.69	96.47	95.29	96.08	94.51	94.51
MRR	0.91	0.95	0.90	0.91	0.94	0.88	0.93	0.95	0.89	0.93	0.93	0.86
Precision	83.40	91.63	87.84	84.71	90.33	87.32	84.71	91.90	87.06	84.71	89.67	86.27
F1	0.89	0.94	0.91	0.90	0.93	0.91	0.90	0.94	0.91	0.90	0.92	0.90
Top-5												
Recall	94.90	95.69	94.51	97.25	96.08	96.47	98.82	96.89	94.90	97.65	95.29	95.29
MRR	0.91	0.94	0.90	0.92	0.94	0.88	0.94	0.95	0.88	0.93	0.93	0.87
Precision	79.45	89.33	88.24	80.31	89.57	87.69	74.27	85.57	88.39	80.16	89.65	88.00
F1	0.86	0.92	0.91	0.88	0.93	0.92	0.85	0.91	0.92	0.88	0.92	0.92

Table 14: Ablation Studies on the NFCORPUS dataset for Top-1, Top-3, and Top-5 retrieval.

	Fixed-Length			Meta-Chunking-MSP			Meta-Chunking-PPL			LLM-Chunk		
Metric	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**
Top-1												
Recall	85.44	94.94	92.41	91.77	94.30	87.97	90.51	94.94	83.54	90.51	93.04	85.44
MRR	0.85	0.95	0.92	0.92	0.94	0.88	0.91	0.95	0.84	0.91	0.93	0.85
Precision	85.44	94.94	92.41	91.77	94.30	87.97	90.51	94.94	83.54	90.51	93.04	85.44
F1	0.85	0.95	0.92	0.92	0.94	0.88	0.91	0.95	0.84	0.91	0.93	0.85
Top-3												
Recall	86.71	94.30	96.84	95.57	94.94	98.10	94.94	97.47	97.47	94.30	95.57	96.20
MRR	0.85	0.94	0.95	0.94	0.95	0.94	0.92	0.96	0.92	0.92	0.94	0.91
Precision	77.22	90.72	93.25	85.65	90.72	92.19	85.02	90.51	89.87	83.97	90.30	89.87
F1	0.82	0.92	0.95	0.90	0.93	0.95	0.90	0.94	0.94	0.89	0.93	0.93
Top-5												
Recall	90.51	96.20	96.20	95.57	96.84	98.73	95.57	98.73	98.10	92.41	96.20	96.84
MRR	0.88	0.94	0.93	0.93	0.95	0.94	0.93	0.96	0.93	0.91	0.94	0.91
Precision	75.95	88.86	92.91	79.49	89.49	92.66	75.32	86.84	91.90	75.06	87.97	90.63
F1	0.83	0.92	0.95	0.87	0.93	0.96	0.84	0.92	0.95	0.83	0.92	0.94

Table 15: Ablation Studies on the HotpotQA dataset for Top-1, Top-3, and Top-5 retrieval.

	Fixed-Length			Meta-Chunking-MSP			Meta-Chunking-PPL			LLM-Chunk		
Metric	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**
Top-1												
Recall	90.36	90.58	88.79	92.15	93.05	86.55	93.50	93.72	81.61	91.70	93.27	82.74
MRR	0.90	0.91	0.89	0.92	0.93	0.87	0.94	0.94	0.82	0.92	0.93	0.83
Precision	90.36	90.58	88.79	92.15	93.05	86.55	93.50	93.72	81.61	91.70	93.27	82.74
F1	0.90	0.91	0.89	0.92	0.93	0.87	0.94	0.94	0.82	0.92	0.93	0.83
Top-3												
Recall	94.84	91.70	90.80	95.29	93.95	94.62	95.96	94.17	95.07	94.84	94.17	92.38
MRR	0.92	0.91	0.88	0.94	0.93	0.89	0.95	0.94	0.89	0.93	0.94	0.86
Precision	85.87	88.19	86.62	87.97	89.91	87.74	88.71	90.51	87.89	87.29	89.91	84.53
F1	0.90	0.90	0.89	0.91	0.92	0.91	0.92	0.92	0.91	0.91	0.92	0.88
Top-5												
Recall	94.62	92.15	91.70	96.41	95.07	94.62	96.19	94.17	94.84	94.17	95.29	94.17
MRR	0.92	0.91	0.88	0.94	0.94	0.89	0.94	0.94	0.88	0.92	0.94	0.86
Precision	82.20	86.68	87.04	84.71	88.97	87.85	84.17	88.21	88.30	81.70	88.52	87.09
F1	0.88	0.89	0.89	0.90	0.92	0.91	0.90	0.91	0.91	0.88	0.92	0.90

Table 16: Ablation Studies on the SCI-DOCS dataset for Top-1, Top-3, and Top-5 retrieval.

	Fixed-Length			Meta-Chunking-MSP			Meta-Chunking-PPL			LLM-Chunk		
Metric	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**
Top-1												
Recall	79.05	83.24	86.31	-	-	-	-	-	-	84.64	87.71	87.15
MRR	0.79	0.83	0.86	-	-	-	-	-	-	0.85	0.88	0.87
Precision	79.05	83.24	86.31	-	-	-	-	-	-	84.64	87.71	87.15
F1	0.79	0.83	0.86	-	-	-	-	-	-	0.85	0.88	0.87
Top-3												
Recall	90.78	93.58	90.78	-	-	-	-	-	-	94.97	95.25	93.02
MRR	0.84	0.88	0.86	-	-	-	-	-	-	0.89	0.91	0.88
Precision	33.89	35.47	83.71	-	-	-	-	-	-	35.85	37.06	85.85
F1	0.49	0.51	0.87	-	-	-	-	-	-	0.52	0.53	0.89
Top-5												
Recall	93.30	94.69	93.02	-	-	-	-	-	-	97.49	98.04	94.97
MRR	0.85	0.88	0.86	-	-	-	-	-	-	0.90	0.92	0.86
Precision	22.12	23.46	83.80	-	-	-	-	-	-	23.63	24.13	84.41
F1	0.36	0.38	0.88	-	-	-	-	-	-	0.38	0.39	0.89

Table 17: Ablation Studies on the SQUAD dataset for Top-1, Top-3, and Top-5 retrieval.

	Fixed-Length			Meta-Chunking-MSP			Meta-Chunking-PPL			LLM-Chunk		
Metric	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**	**Base**	**MM**	**Momoka**
Top-1												
Recall	45.56	60.62	75.29	59.85	67.18	79.15	68.34	74.13	78.38	64.86	71.04	81.85
MRR	0.10	0.10	0.75	0.13	0.13	0.79	0.14	0.14	0.78	0.11	0.14	0.82
Precision	45.56	60.62	75.29	59.85	67.18	79.15	68.34	74.13	78.38	64.86	71.04	81.85
F1	0.46	0.61	0.75	0.60	0.67	0.79	0.68	0.74	0.78	0.65	0.71	0.82
Top-3												
Recall	67.18	79.92	86.10	82.24	84.94	87.64	86.10	90.73	88.80	87.26	86.49	86.87
MRR	0.12	0.12	0.81	0.15	0.15	0.83	0.16	0.16	0.81	0.14	0.15	0.83
Precision	32.05	51.99	77.35	41.44	53.80	80.57	45.05	54.05	79.41	46.46	53.54	81.21
F1	0.43	0.63	0.81	0.55	0.66	0.84	0.59	0.68	0.84	0.61	0.66	0.84
Top-5												
Recall	77.22	85.71	87.64	89.58	91.51	88.04	91.12	92.28	88.03	92.66	91.51	87.64
MRR	0.13	0.12	0.81	0.16	0.16	0.82	0.17	0.16	0.78	0.15	0.16	0.82
Precision	26.64	46.95	77.07	33.44	46.64	78.69	32.28	41.47	76.29	33.36	42.08	79.00
F1	0.40	0.61	0.82	0.49	0.62	0.83	0.48	0.57	0.82	0.49	0.58	0.83

Table 18: Ablation Studies on the TriviaQA dataset for Top-1, Top-3, and Top-5 retrieval.

Metric	Iter 20 / Roll 3	Iter 20 / Roll 5	Iter 50 / Roll 1	Iter 50 / Roll 3	Iter 50 / Roll 5	Iter 100 / Roll 1	Iter 100 / Roll 3
Top-1							
Recall	21.01	20.85	20.85	20.80	21.01	22.87	22.66
MRR	0.75	0.74	0.75	0.74	0.75	0.77	0.76
Precision	74.94	74.06	75.08	74.47	75.38	76.81	76.09
F1	0.33	0.33	0.33	0.33	0.33	0.35	0.35
Top-3							
Recall	25.76	25.41	25.32	25.60	25.69	30.00	28.85
MRR	0.81	0.79	0.79	0.80	0.79	0.82	0.80
Precision	77.42	76.77	77.21	77.48	77.38	79.62	77.35
F1	0.39	0.38	0.39	0.38	0.39	0.44	0.42
Top-5							
Recall	33.48	32.87	32.31	32.05	33.06	35.84	34.78
MRR	0.79	0.78	0.78	0.78	0.79	0.83	0.83
Precision	70.91	69.77	70.66	69.81	71.25	78.70	77.36
F1	0.45	0.45	0.44	0.44	0.45	0.49	0.48

Table 19: Sensitivity Analysis of Critical Hyper-parameters on the Dragonball dataset for Top-1, Top-3, and Top-5 retrieval.

{"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 6: 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 7: 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 8: Detailed informations of NFCORPUS dataset.

{"_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 9: 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 10: 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 11: Detailed informations of TriviaQA dataset.

Top-k	Method	Recall	Precision	MRR	F1
Top-1	w/	22.96	77.11	0.77	0.35
	w/o	22.66	76.09	0.76	0.35
Top-3	w/	28.94	79.55	0.83	0.42
	w/o	28.85	77.35	0.80	0.42
Top-5	w/	34.66	78.25	0.83	0.48
	w/o	34.78	77.36	0.83	0.48

Table 20: Analysis of MCTS Reward on the Dragonball dataset for Top-1, Top-3, and Top-5 retrieval.

Top-k	Method	Recall	Precision	MRR	F1
Top-1	w/	80.78	80.78	0.81	0.81
	w/o	79.61	79.61	0.80	0.80
Top-3	w/	94.90	86.54	0.87	0.91
	w/o	94.51	86.27	0.86	0.90
Top-5	w/	94.51	87.14	0.85	0.91
	w/o	95.29	88.00	0.87	0.92

Table 21: Analysis of MCTS Reward on the NF-CORUPS dataset for Top-1, Top-3, and Top-5 retrieval.

Top-k	Method	Recall	Precision	MRR	F1
Top-1	w/	82.74	82.74	0.83	0.83
	w/o	82.74	82.74	0.83	0.83
Top-3	w/	93.50	85.13	0.87	0.89
	w/o	92.38	84.53	0.86	0.88
Top-5	w/	94.39	87.31	0.87	0.91
	w/o	94.17	87.09	0.86	0.90

Table 22: Analysis of MCTS Reward on the SCI-DOCS dataset for Top-1, Top-3, and Top-5 retrieval.

Top-k	Method	Recall	Precision	MRR	F1
Top-1	w/	81.85	81.85	0.82	0.82
	w/o	81.85	81.85	0.82	0.82
Top-3	w/	88.42	81.98	0.84	0.85
	w/o	86.87	81.21	0.83	0.84
Top-5	w/	259.46	242.59	2.49	2.51
	w/o	256.36	242.06	2.47	2.49

Table 23: Analysis of MCTS Reward on the TriviaQA dataset for Top-1, Top-3, and Top-5 retrieval.