SAKI-RAG: Mitigating Context Fragmentation in Long-Document RAG via Sentence-level Attention Knowledge Integration

Wenyu Tao[1], Xiaofen Xing[1]*, Zeliang Li[1], Xiangmin Xu[2,1]
[1]School of EE., South China University of Technology, Guangzhou, China
[2]Foshan University, Foshan, China
eetaowenyu@mail.scut.edu.cn, zeliang0li@163.com, {xfxing, xmxu}@scut.edu.cn

Abstract

Traditional Retrieval-Augmented Generation (RAG) frameworks often segment documents into larger chunks to preserve contextual coherence, inadvertently introducing redundant noise. Recent advanced RAG frameworks have shifted toward finer-grained chunking to improve precision. However, in long-document scenarios, such chunking methods lead to fragmented contexts, isolated chunk semantics, and broken inter-chunk relationships, making cross-paragraph retrieval particularly challenging. To address this challenge, maintaining granular chunks while recovering their intrinsic semantic connections, we propose SAKI-RAG (Sentence-level Attention Knowledge Integration Retrieval-Augmented Generation). Our framework introduces two core components: (1) the SentenceAttnLinker, which constructs a semantically enriched knowledge repository by modeling inter-sentence attention relationships, and (2) the Dual-Axis Retriever, which is designed to expand and filter the candidate chunks from the dual dimensions of semantic similarity and contextual relevance. Experimental results across four datasets—Dragonball, SQUAD, NFCORPUS, and SCI-DOCS demonstrate that SAKI-RAG achieves better recall and precision compared to other RAG frameworks in long-document retrieval scenarios, while also exhibiting higher information efficiency.

1 Introduction

RAG, initially proposed by Lewis et al. (2021), was designed to enhance LLMs' performance in domain-specific tasks and mitigate hallucinations (Augenstein et al., 2023; Huang et al., 2025). Its core mechanism involves dynamically retrieving relevant text chunks from external knowledge bases to supplement LLMs, thereby overcoming the limitations of static training data dependency.

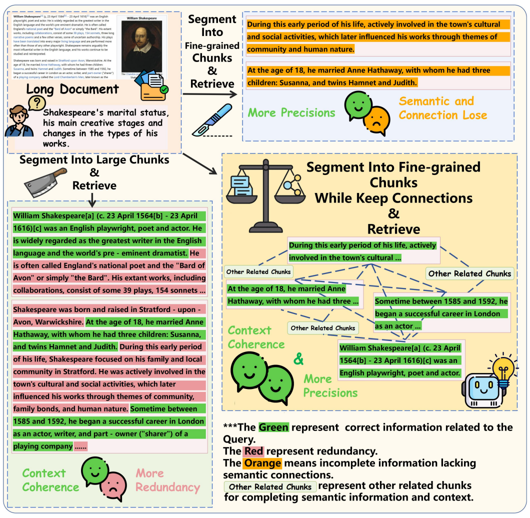

Figure 1: In long-document cross-paragraph retrieval, **Large Chunks** ensure context coherence but add redundancy. **Fine-grained Chunks** offer more precision but risk semantic and informational loss. The solution is to balance both, keeping chunks fine-grained yet interconnected.

As LLMs increasingly handle complex tasks involving long documents, directly inputting entire documents as context becomes impractical (Jin et al., 2024). Consequently, RAG techniques are employed to split long documents into chunks and precisely recall relevant ones for high - quality answers. Traditional frameworks like Naive RAG use fixed length or regularized document splitting, storing chunks in local vector databases via embedding models and retrieving them through methods like BM25 (Robertson et al., 1996) or cosine similarity (Zhang et al., 2020). Recent RAG frameworks have evolved with various innovative approaches. For instance, Late-Chunking (Günther et al., 2024) adopts an "embedding then chunking" strategy, allowing each chunk to retain contextual information in its embeddings. Meta-Chunking (Zhao et al., 2024) dynamically determines chunk sizes by using LLMs with Margin Sampling (MSP) Chunk-

*Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 1196–1214
November 4-9, 2025 ©2025 Association for Computational Linguistics

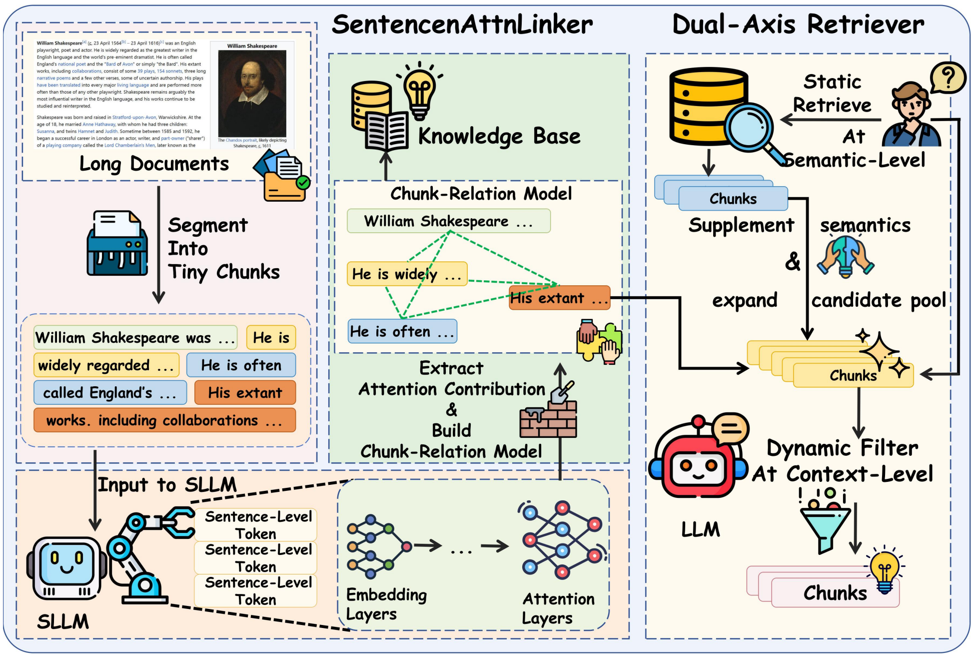

Figure 2: Framework of SAKI-RAG.

ing or Perplexity (PPL) Chunking. Dense X Retrieval (DXR) (Chen et al., 2024) decomposes text into finer units called propositions. The RAPTOR framework (Sarthi et al., 2024) treats each chunk as a leaf node and constructs a tree-structured knowledge base through bottom-up soft clustering and summarization. Frameworks like GraphRAG (Edge et al., 2024), LightRAG (Guo et al., 2025), and nano-GraphRAG (gusye1234, 2024) extract entities from chunks and connect them using a graph structure. However, these methods struggle with long documents. Larger chunks provide more information but lack precision, while smaller chunks offer precision but lose information and connections, as shown in Figure 1.

To address the aforementioned issues, we propose SAKI-RAG, which consists of two components: SentenceAttnLinker and Dual-Axis Retrieval. In the SentenceAttnLinker component, we adopt the SLLM proposed by An et al. (2024) as a critical module. The SLLM operates at the sentence level rather than the token level, implemented through a Sentence Variational Autoencoder (Sentence-VAE) integrated by reconstructing the input and output layers of a standard LLM. After segmenting the entire document into fine-grained chunks, we feed them collectively into the SLLM. Since the SLLM processes text at the sentence level, its capacity to handle long-document content is significantly enhanced. We then compute attention contributions between sentences using the self-attention layer weights of the SLLM, thereby modeling inter-sentence correlations. In the Dual-Axis Retriever component, we retrieve and filter chunks through two dimensions. Initially, we perform retrieval at the semantic similarity dimension using static methods to swiftly identify relevant chunks. Then, we expand the candidate pool by incorporating chunks relevant at the contextual relevance dimension, as determined by the SentenceAttnLinker phase. Meanwhile, we bring in the LLM's deep semantic reasoning capability to dynamically filter chunks according to the user's question. This approach alleviates the negative optimization issues in reranking caused by the semantic deficiencies in fine-grained chunks.

To demonstrate the superiority of our framework, we conducted experiments on the Dragonball (Zhu et al., 2025), SQUAD (Rajpurkar et al., 2016), NFCORPUS (Boteva et al., 2016), and SCI-DOCS (Cohan et al., 2020) datasets which are filtered. The evaluation metrics used were Recall@k (Musgrave et al., 2020), Precision@k, and Information-Efficiency@k(IE@k). The experimental results indicate that, compared to other RAG frameworks, our proposed framework achieves better performance in long-document retrieval scenarios.

Main contributions of this paper are as follows:

(1)We present SentenceAttnLinker, which leverages the attention contributions of sentence-level tokens from SLLM to build a chunk-relation model. This effectively avoids the gap between word-level tokens and sentence-level semantics.

(2)We propose Dual-Axis Retriever, which combines static and dynamic methods to retrieve and filter chunks across two dimensions—semantic similarity and contextual relevance—according to users' questions.

(3)Our framework delivers excellent performance on the Dragonball, SQUAD, NFCORPUS, and SCI-DOCS datasets. It demonstrates remarkable recall and precision along with superior information efficiency.

2 Related Work

As LLM advances, their comprehension and generation abilities have improved. Yet, they still make factual errors in specialized domains (Zhao et al., 2025), necessitating the inclusion of relevant information as context alongside questions. However, with growing complexity of tasks and the increasing prevalence of long documents, using an entire long document as context is impractical, leading to issues like model input limitations and loss of attention focus.

Langchain[1] (Chase, 2024) provides various traditional chunking strategies, such as RecursiveCharacterTextSplitter and Character-TextSplitter. These methods, which split documents based on fixed lengths or rules, are better suited for scenarios where precision and context coherence are not critical. They struggle with complex questions in long-document settings.

Late-Chunking, a popular RAG framework, adopts an "embed-then-chunk" strategy. This approach maintains chunk fine-grained while incorporating context into each chunk's embedding vector through average pooling. Nevertheless, long documents, with their excessive tokens, often exceed the embedding model's input limit. This requires batch processing, which can lead to context fragmentation. Additionally, the high volume of tokens may dilute the informational density of the embeddings.

Meta-Chunking integrates LLMs with MSP Chunking and PPL Chunking to dynamically control chunk size for better context coherence. However, when relevant information is dispersed across the text, this method may truncate necessary details.

[1]https://www.langchain.com/

Dense X Retrieval focuses on decomposing text into fine-grained propositions, each encapsulating a unique factual element. While innovative, DXR may struggle to capture the complex relationships and overall semantics within long documents, as it processes each proposition independently.

To address these challenges, some RAG frameworks are exploring ways to link chunks. RAPTOR, for instance, constructs a tree structure from chunks as leaf nodes through soft clustering and summarization. However, this approach treats all chunks within a cluster as equivalent, and smaller chunks can result in weaker, more easily confused semantic information.

Frameworks such as GraphRAG, LightRAG and nano-GraphRAG organize chunks into a graph structure. However, large chunks may introduce redundancy, causing the LLM to become "lost in the middle (Liu et al., 2023)," while small chunks might lack key entity information, thereby affecting the quality of the generated graph.

3 SAKI-RAG

In this section, we will introduce in Section 3.1 how SentenceAttnLinker utilizes SLLM to calculate the attention contributions between chunks for chunk-relation model, as well as how Dual-Axis Retriever performs static and dynamic retrieval and filtering in the knowledge base with sentence-level relevance metadata built by SentenceAttnLinker to obtain the most relevant chunks. The framework is shown in Figure 2.

3.1 SentenceAttnLinker

Most LLMs primarily use word-level tokens, focusing on word-to-word attention relationships and employing self-attention mechanisms (Vaswani et al., 2023) to capture complex word dependencies in text sequences. Inspired by this, we aim to apply attention mechanisms to discovering relationships between chunks. However, employing popular embedding models like BGE-M3 (Chen et al., 2023) – originally designed for word-level semantic interactions through attention mechanisms – to establish sentence-level relationships introduces gap. The SLLM proposed by An et al. (2024) offers a useful tool to bridge this gap. In SLLM, training and encoding are sentence-level-token-based, allowing long documents to be processed in one go. Since it uses sentence-level rather than word-level tokens, input token limits are rarely exceeded. More-over, SLLM's attention layers are better suited for sentence-level tokens processing. In SentenceAttnLinker, we extract certain layers from SLLM as a core component to build a Chunk-Relation Model and local knowledge base.

After cleaning the long document, we use a regularization tool to quickly chunk the long document into fine-grained chunks. The resulting collection of sentences is denoted as $S = \{s_1, s_2, ..., s_n\}$. We then employ the SentenceVAE encoder to generate sentence vectors $\{\Omega_i\}$ which is determined by the following formula:

$$\Omega_i = SentenceVAE - Encoder(s_i), \quad (1)$$

where $\Omega_i \in R^d$, d is the hidden layer dimension.

After adding positional encoding to the vector sequence $\{\Omega_i\}$ to create initial hidden states and inputting them into the SLLM, we compute, for each layer l of the LLM and each attention head h, the query matrix $\mathbf{Q}^{(l,h)}$ and key matrix $\mathbf{K}^{(l,h)}$. The attention weight matrix is generated via Softmax:

$$\text{Attn}^{(l,h)} = \text{softmax}\left(\frac{\mathbf{Q}^{(l,h)}(\mathbf{K}^{(l,h)})^\top}{\sqrt{d}}\right) \in \mathbb{R}^{n\times n} \quad (2)$$

Ultimately, we obtain attention contribution matrix $A \in \mathbb{R}^{n\times n}$, which serves as the chunk-relation model. Here, A_{ij} represents the attention contribution of sentence s_i to s_j:

$$A_{ij} = \frac{1}{L \cdot H} \sum_{l=1}^{L} \sum_{h=1}^{H} \text{Attn}_{ij}^{(l,h)}, \quad (3)$$

where L is the number of LLM layers, and H is the number of attention heads per layer.

For each sentence s_i, extract the corresponding attention contribution row A_i, sort related chunks in descending order to get $\{s_{i_1}, s_{i_2}, s_{i_3}, \dots\}$, and record the weights. The final storage structure is:

$$\begin{aligned} Metadata[s_i] = [&(s_{i_1}, A_{i,i_1}), \\ &(s_{i_2}, A_{i,i_2}), \\ &(s_{i_3}, A_{i,i_3}), \dots] \end{aligned} \quad (4)$$

The sentence vectors $\{\Omega_i\}$ are stored in a vector database along with the above metadata, forming an efficient semantic index for retrieval.

3.2 Dual-Axis Retriever

Traditional RAG often directly uses BM25, cosine similarity retrieval, and other retrieval strategies to retrieve chunks, and then screens them through a Rerank Model to obtain the final Top-k chunks. However, chunk size poses a problem. Large chunks, while including more information, bring in redundancy that dilutes or overshadows key details. Fine-grained chunks, though offering higher precision, lose contextual links and semantic information, like subject terms. Thus, searching and filtering solely based on semantic similarity may not identify the chunks most relevant to the user's question.

Popular RAG frameworks use LLMs to determine chunk relevance to the question after retrieving chunks. But fine-grained chunks, often missing subjects and other key information, make it hard for LLMs to accurately assess their relevance.

To address these issues, we propose a Dual-Axis Retriever that combines static retrieval and dynamic filtering. This ensures retrieved chunks have both semantic similarity and contextual relevance to the user's question.

Algorithm 1: Dual-Axis Retriever

Input: Query Q, Vector DB V, LLM M, Reranker R, Top_k
Output: Retrieved chunks F

$C_{\text{init}} \leftarrow V.\text{search}(Q, Top\_k)$ `// Static semantic retrieve`
$C_{\text{filt}} \leftarrow \emptyset$
for $c \in C_{\text{init}}$ **do**
  $R_c \leftarrow \text{parse}(c.\text{meta}[\texttt{"related"}])$ `// Get related chunks`
  for $r \in R_c$ **do**
    $k_c \leftarrow c.\text{content} \oplus r$
    $p \leftarrow$ "`Determine relevance: // Knowledge:` k_c `// Question:` Q `// Output: 1/0`"
    if $M(p) = 1$ **then**
      $C_{\text{filt}} \leftarrow C_{\text{filt}} \cup \{k_c\}$
    end if
  end for
end for
$F \leftarrow R.\text{rerank}(C_{\text{filt}}, Q, Top\_k)$ `// Context-aware ranking`
return F

Description: V.search($\cdot$) refers to using retrieve methods such as BM25 and cosine similarity to retrieve chunks.

Given a user query q, it is embedded into a vec-

Dataset	Ave_Doc_Length
Dragonball	11436
SQUAD	2303
NFCORPUS	3267
SCI-DOCS	7955

Table 1: Average Document Length of Each Dataset

tor. Cosine similarity is used to retrieve an initial candidate set C_{init} from the vector library created by the SentenceAttnLinker component:

$$C_{\text{init}} = \{s_i\}, \quad \|C_{\text{init}}\| = Top\_k \tag{5}$$

For each candidate sentence s_i in C_{init}, perform context expansion and relevance determination. Extract the associated sentence set $R_i = \{r_{i1}, r_{i2}, \ldots, r_{i_{Top-k}}\}$, which are sorted by self-attention weights from the metadata, and generate a context-enhanced candidate block $k_i = s_i \oplus R_i$, where $\oplus$ denotes string concatenation.

Then input k_i and uses' question q into a pre-trained large language model api, such as Qwen-max (Bai et al., 2023) which has strong comprehension ability, to judge their relevance via a binary classification task:

$$\text{Score}_{\text{rel}}(k_i, q) = \mathbb{I}\left(\text{LLM}([k_i; q]) \rightarrow \text{“1”}\right), \tag{6}$$

where $\mathbb{I}(\cdot)$ is an indicator function that retains candidates with $\text{Score}_{\text{rel}} = 1$, forming the filtered set C_{filtered}. For detailed prompt information, please refer to Appendix A.1.

For the candidates in C_{filtered}, use a reranker to calculate the final relevance score:

$$Score_{\text{final}}(k_i, q) = Reranker(k_i, q) \tag{7}$$

Candidates are finally ranked and recalled based on $\text{Score}_{\text{final}}$. In Algorithm 1, we show this retrieval strategy.

4 Experiments

Datasets. Experiments were conducted on four datasets: Dragonball, SQUAD, NFCORPUS, and SCI-DOCS, filtered by document length. The average document length for each dataset is in Table 1. Only the Finance subset of Dragonball was used, as other subsets contain structured content like legal judgments and medical records, not coherent text.

Embedding Model and Reranker. The framework we propose and the baseline for comparison don't rely on specific embedding models, and changing embedding models doesn't significantly affect functionality or ranking. Thus, in all experiments, we used the BGE-M3 embedding model, which performs well across languages and domains. 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, with all model parameters being the default parameters of the BCERerank function in the BCEmbedding repository [2].

LLM. In parts involving calling pre-trained LLMs for entities extracting, filtering and answer generation tasks, we use the LangChain-based Tongyi model interface to call Qwen-max, a model with strong understanding and performance, with all parameters at their default settings. For Meta-Chunking, we deploy Qwen-2-1.5B locally for text chunking. In the SAKI-RAG framework, we use a 1.3B-parameter SLLM model with default settings from the SentenceVAE repository[3].

Chunks Size. To keep experimental variables consistent, for frameworks requiring custom chunk input like SAKI-RAG, LightRAG and RAPTOR, we use a regularization tool to split documents into chunks of two sentences each. For frameworks needing an expected chunk size, such as Meta-Chunking, we set $target\_size$ to 50, matching the earlier average chunk length. Other frameworks are left at their default settings.

Metrics. For retrieve evaluation metrics, we chose $Recall$, $Precision$, and IE. For generation evaluation, we use $ROUGE-L$ (Lin, 2004) and $METEOR$ (Banerjee and Lavie, 2005). $IE@k$ measures the framework's ability to retrieve effective information in search tasks, is calculated as Formula 8 .

$$IE@k = Recall@k \times Precision@k \tag{8}$$

The final metric score is computed using Formula 9.

$$Metric = Metric@1 + Metric@3 + Metric@5 \tag{9}$$

4.1 Comparative Experiments

In terms of retrieval performance, we compare our proposed SAKI-RAG with popular RAG

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

[3] https://github.com/BestAnHongjun/SentenceVAE

Methods	Dragonball			SQUAD			NFCORPUS			SCI-DOCS		
	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑
Late-Chunking	2.24	3.51	0.02	70.67	34.58	7.58	12.95	5.47	0.21	2.01	1.02	0.01
RAPTOR	96.14	125.07	35.79	/	/	/	283.14	257.86	243.09	293.05	279.36	272.81
Meta-Chunking-PPL	128.21	133.95	50.96	254.19	133.04	110.37	287.06	245.00	233.76	65.02	54.75	11.85
Meta-Chunking-MSP	97.86	125.68	36.59	257.27	133.61	111.53	283.92	252.42	238.41	288.72	264.91	264.60
Dense X Retrieval	7.46	14.16	0.32	204.19	104.21	67.07	279.78	255.96	238.48	287.22	264.60	253.12
SAKI-RAG	106.09	235.32	83.98	277.40	282.06	260.95	262.35	285.06	249.28	274.89	292.52	268.04

Table 2: **Comparative Experiments on Retrieval:** Due to the presence of sensitive or unsafe content in the original documents of the SQUAD, LLMs cannot be used to build tree structures. In the table, we abbreviate the metrics, where **Rec.**, **Pre.**, and **IE** stand for Recall, Precision, and Information Efficiency respectively.

Methods	ROUGE-L↑	METEOR↑
LightRAG	0.2865	0.2852
SAKI-RAG	0.3122	0.3254

Table 3: **Comparative Experiments on Generation:** Only the Dragonball dataset provides human-annotated detailed answers, so we only conduct generation quality experiments on it.

frameworks like Late-Chunking, RAPTOR, Meta-Chunking PPL, Meta-Chunking MSP, and Dense X Retrieval. For generation quality, we contrast it with LightRAG, an enhanced customization-wise version of GraphRAG, in mix mode with the $response\_type$ set to output answers in a single paragraph without sources and references. Retrieval results are in Table 2, and generation quality results are in Table 3.

In this subsection's experiments on retrieval quality, we compare SAKI-RAG with popular recall-focused RAG frameworks: Late-Chunking, RAPTOR, Meta-Chunking PPL, Meta-Chunking MSP, and Dense X Retrieval. In the table, the top two frameworks' scores are highlighted in blue, with darker shades for the first place and lighter for the second. Recall scores show SAKI-RAG has decent results, though not the highest. However, some frameworks that segment documents into larger chunks may have artificially inflated Recall metrics due to chunks containing more content. This is why we include Precision and IE metrics. Precision reflects the accuracy of recalls, and IE indicates the effectiveness of the recalled information. SAKI-RAG excels in Precision, often achieving the best results. More importantly, it also performs well in IE. This means SAKI-RAG maintains high accuracy and information effectiveness while achieving good recall performance.

In the four datasets of the comparative experiments, the Dragonball dataset comprises numerous cross-paragraph retrieval problems, including summarization and multi-hop questions. In contrast, the SQUAD, NFCORPUS, and SCI-DOCS datasets consist of factual questions involving single entities. The experimental results indicate that SAKI-RAG has achieved the best Precision metric scores across all dataset experiments and has also secured top positions in IE metric in most of the dataset experiments. This demonstrates that SAKI-RAG can deliver superior performance when handling cross-paragraph retrieval problems in long-document contexts while maintaining decent performance on conventional factual questions. Despite not achieving the highest Recall scores in some datasets due to the influence of chunk size on answer coverage, SAKI-RAG, which adopts fine-grained chunks, still attains respectable scores. For more information about results of experiments, please refer to the Appendix A.4.

To explore where SAKI-RAG performs best, we divide the Dragonball dataset by question type into subsets and run comparative experiments. As shown in Table 4, SAKI-RAG achieves the highest Precision and IE scores across all subset experiments, particularly excelling in Non-Factual questions. Compared to previous experiments, SAKI-RAG not only performs well in typical retrieval tasks but also shows superior performance in non-factual questions like Multi-hop Reasoning and Summary Questions. This demonstrates SAKI-RAG's better handling of cross-paragraph retrieval in long document.

In the generation quality experiments of this subsection, we compare SAKI-RAG with LightRAG, a framework focused on answer generation. We highlight better results in the table. The results show that SAKI-RAG can achieve scores comparable to LightRAG with a simpler framework.

Methods	Dragonball-Hop			Dragonball-Summary			Dragonball-Non-Factual		
	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑
RAPTOR	137.59	159.72	65.95	55.77	107.55	18.50	88.57	136.89	36.35
Meta-Chunking-PPL	178.59	161.14	86.75	82.18	120.39	30.30	120.82	143.31	51.90
Meta-Chunking-MSP	146.91	162.60	71.83	51.31	101.13	15.77	89.62	135.70	36.42
Dense X Retrieval	10.84	19.70	0.65	3.75	12.18	0.13	6.58	16.41	0.33
Late-Chunking	2.52	3.58	0.03	1.56	4.24	0.02	1.94	3.86	0.02
SAKI-RAG	144.46	276.06	133.46	63.82	171.12	37.67	96.34	230.02	74.80

Table 4: **Comparative Experiments of Different Query Types on Dragonball:** The Dragonball dataset divides questions into subtypes like *Multi-hop Reasoning Question*, *Summary Question*, and *Factual Question*. We conduct further refined experiments on this dataset to explore which question type SAKI-RAG performs great on. In the table, **Dragonball-Hop**, **Dragonball-Summary**, and **Dragonball-Non-Factual** respectively represent experiments conducted exclusively on Multi-hop Reasoning Questions, Summary Questions, and question types other than Factual Questions.

Methods	Dragonball			SQUAD			NFCORPUS			SCI-DOCS		
	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑	Rec.↑	Pre.↑	IE↑
Naive	92.09	128.61	34.75	273.81	146.03	130.91	288.63	144.15	135.67	283.18	264.20	249.23
Naive+SAL	105.84	227.30	81.47	277.93	265.87	246.47	285.10	282.04	268.07	282.73	280.81	264.65
SAKI	106.09	235.32	83.98	277.40	282.06	260.95	262.35	285.06	249.28	274.89	292.52	268.04

Table 5: **Ablation Studies:** In the table, **"Naive"** stands for Naive RAG, which maintains a consistent chunk size, directly embeds chunks into vector space, and retrieves chunks via cosine similarity. **"SAL"** refers to using SentenceAttnLinker for chunking and Embedding while still employing cosine similarity for retrieval. **"SAKI"** denotes SAKI-RAG, which incorporates the Dual-Axis Retriever strategy in addition to SentenceAttnLinker.

In Appendix A.6, we present additional experimental results, such as evaluations on the HotpotQA and TriviaQA datasets. We also include more retrieval performance metrics including F1, EIR, MRR, and latency, as well as generation quality indicators such as Relevant, Irrelevant, and Wrong. In Appendix A.7, we provide statistical validation experiments to demonstrate that our results achieve statistically significant wins.

4.2 Ablation Studies

SAKI-RAG is built on the SentenceAttnLinker chunking method and incorporates the Dual-Axis Retriever strategy. To verify the effectiveness of each component in the framework, ablation experiments are conducted on the datasets in this section. The results are shown in Table 5.

In the ablation study of the SAKI-RAG framework, we thoroughly analyze its components, especially focusing on the performance differences across various datasets. The experimental results show that on the Dragonball and SQUAD datasets, as the components were gradually improved, the Recall, Precision, and IE metrics show a positive upward trend, with Precision and IE being particularly prominent. On the NFCORPUS and SCI-DOCS datasets, although Precision and IE metrics show an upward trend, the Recall metric decline.

In the Dragonball and SQUAD datasets, our framework demonstrated effective handling of cross-paragraph retrieval problems. This is attributed to its ability to integrate multiple relevant paragraphs in the context of long documents. The SentenceAttnLinker is able to capture sentence-to-sentence relationships, and the Dual-Axis Retriever further enhance retrieval accuracy through its dual-dimensional filtering mechanism, leading to the framework's superior performance on these datasets. However, in the NFCORPUS and SCI-DOCS datasets, the type of questions and the characteristics of the dataset content become key factors affecting the metric performance. For detailed dataset information, please refer to Appendix A.5. For more information about results of experiments, please refer to the Appendix A.3

Unlike Dragonball, which involve cross-paragraph retrieval problems with multiple entities, the NFCORPUS and SCI-DOCS datasets consist of factual questions involving only a single entity. In the SAL, on the one hand, chunk concatenation leads to longer chunk content, which dilutes the original semantic information to some extent. As a result, after reranking based on the user's query, the correct chunks rank lower. On the other hand, chunk concatenation may introduces semantic information relevant to the user's question. However, the chunks themselves are incorrect answers, causing the reranked incorrect chunks to rise in ranking and ultimately leading to a decline in the Recall metric of SAL.

Compared to others, NFCORPUS and SCI-DOCS are more specialized datasets. For instance, NFCORPUS is a medical-information dataset. The LLM filtering mechanism introduced in SAKI may have certain limitations in processing the semantic information of professional academic terms. The LLM may have deviations in understanding domain-specific terminology and complex logical structures in academia, causing some chunks that should have been recalled to fail the screening and thus leading to a decline in the Recall metric. On the other hand, the content expansion caused by chunk concatenation dilutes or obscures some correct key semantic information, resulting in incorrect screening by the LLM.

5 Conclusions

In this paper, we present SAKI-RAG to maintain chunk fine-grained and connections for better long document retrieval. It has two key components: SentenceAttnLinker and Dual-Axis Retriever. SentenceAttnLinker innovatively uses attention mechanisms with SLLM to build a Chunk-Relation Model, uncovering chunk relationships. Dual-Axis Retriever integrates both static retrieval and dynamic filtering strategies, utilizing semantic similarity and contextual relevance to improve the efficiency of chunk selection.

Through comparative, generation, and ablation experiments across four datasets—Dragonball, SQUAD, NFCORPUS, SCI-DOCS, we show SAKI-RAG offers good recall, precision, and information efficiency in long document settings. Also, except for using SLLM, SAKI-RAG doesn't rely on specific embedding models or pre-trained LLMs, involves no extra training, and is widely applicable.

Limitations

During our research, we identified several limitations:

(1)When processing the attention contribution matrix, we didn't distinguish the importance of each layer's contributions and simply averaged them. This might weaken the influence of more critical layers. We plan to explore this issue further in future research to develop more effective matrix construction methods.

(2)Our Chunk-Relation Model, which uncovers relationships between chunks, is limited to chunks within the same document. That is to say, SAKI-RAG is adept at tackling cross-paragraph retrieval, but it might not hold a significant edge when it comes to cross-document issues. However, when calculating the attention contributions between chunks, it is necessary to add position encoding information. If we want to explore the relationships between chunks from different documents, how to add position encoding information and how to determine the order of chunks from different documents will be challenging issues.

Acknowledgments

This work was supported by Guangdong Provincial Key Research and Development Projects (2024B0101040004), Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004), Nansha Key Project under Grant 2022ZD011, Hainan Province Health and Family Planning Commission Joint Innovation Project (WSJK2025QN011).

References

Hongjun An, Yifan Chen, Zhe Sun, and Xuelong Li. 2024. Sentencevae: Enable next-sentence prediction for large language models with faster speed, higher accuracy and longer context.

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.

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, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

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.

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.

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?

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

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.

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

gusye1234. 2024. nano-graphrag: A simple, easy-to-hack GraphRAG implementation. `https://github.com/gusye1234/nano-graphrag`.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag.

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.

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.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check.

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

Stephen E Robertson, Steve Walker, MM Beaulieu, Mike Gatford, and Alison Payne. 1996. Okapi at trec-4. *Nist Special Publication Sp*, pages 73–96.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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.

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, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models.

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.

A Appendix

A.1 Detailed Prompt in Dual-Axis Retriever

In this section, we present the detailed prompt in Dual-Axis Retriever in Figure 3.

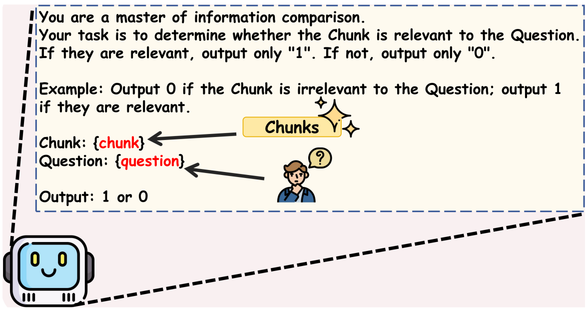

Figure 3: Detailed prompt in Dual-Axis Retriever.

A.2 Chunks Relationship Diagram of SAKI-RAG

In this section, we present an example of the connections between chunks processed by SAKI-RAG. In Figure 4, the pink part represents the original content of the document, the yellow part represents a specific chunk within the document, and the green parts represent chunks related to this yellow chunk. These related chunks are ordered by the magnitude of their attention contributions.

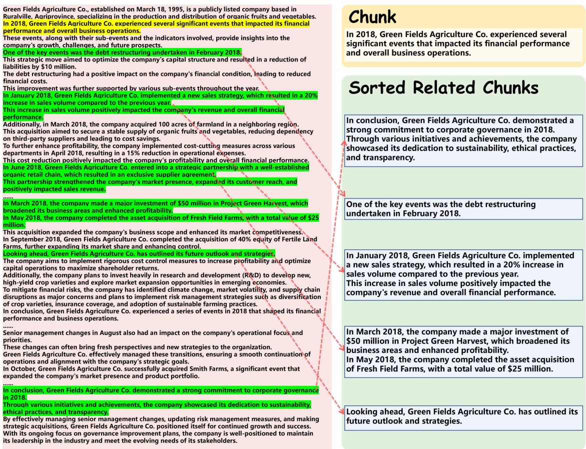

Figure 4: Example of chunks relationship diagram In SAKI-RAG.

A.3 Detailed Information of Comparative Experiments

In this section, we will show more detailed information about comparative experiments on Table 6, 7, 8, 9, 10, 11, 12.

A.4 Detailed Information of Ablation Studies

In this section, we will show more detailed information about Ablation Studies on Table 13, 14, 15, 16.

A.5 Detailed Information of Dataset

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.

A.6 More Comparative Experiments Results

On the original basis, we add retrieval metrics including , average latency time , F1 , MRR , EIR (this metric is proposed by RAGEval and is suitable for the Dragonball dataset constructed by the RAGEval project), etc. At the same time, to further illustrate the superiority of our framework, we add experiments on two new datasets: HotpotQA and TriviaQA . Additionally, we replaced the Qwen-max LLM in our framework with Qwen-Turbo — which offers faster inference speed though slightly weaker reasoning capabilities — to showcase the impact of different LLMs on latency performance.

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 experimental results are presented in Table 17, 18, 19, 20, 21, 22, 23. Due to API updates and the input containing sensitive information, some framework that requires LLM cannot work.

A.7 Statistical Validation Experiments

For retrieval metrics, we employed permutation tests for verification; for generation quality metrics, we adopted stratified sign tests for validation. We use ★ ($p < 0.05$) to mark statistically significant wins. The experimental results are presented in Table 24, 25, 26, 26, 27, 28, 29.

Method	SAKI-RAG	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chuning -MSP	Dense X Retrieval
Top-1						
Rec.	22.14	0.52	19.95	26.75	20.71	1.79
Pre.	74.84	1.83	63.28	68.67	64.50	7.73
IE	16.57	0.01	12.62	18.37	13.36	0.14
Top-3						
Rec.	36.96	0.75	34.28	45.93	34.54	2.66
Pre.	79.76	0.95	35.84	38.08	35.16	3.83
IE	29.48	0.01	12.29	17.49	12.14	0.10
Top-5						
Rec.	46.99	0.97	41.91	55.53	42.61	3.01
Pre.	80.72	0.73	25.95	27.20	26.02	2.60
IE	37.93	0.01	10.88	15.10	11.09	0.08

Table 6: **Detailed information of comparative experiments on Dragonball.**

Method	SAKI-RAG	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chuning -MSP	Dense X Retrieval
Top-1						
Rec.	86.33	16.48	/	80.45	80.17	56.98
Pre.	92.49	16.48	/	80.45	80.17	56.98
IE	79.85	2.72	/	64.72	64.27	32.47
Top-3						
Rec.	95.53	25.70	/	86.59	87.99	71.79
Pre.	95.18	10.61	/	32.31	32.77	28.12
IE	90.93	2.73	/	27.98	28.83	20.19
Top-5						
Rec.	95.54	28.49	/	87.15	89.11	75.42
Pre.	94.39	7.49	/	20.28	20.67	19.11
IE	90.18	2.13	/	17.67	18.42	14.41

Table 7: **Detailed information of comparative experiments on SQUAD.**

Method	SAKI-RAG	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chuning -MSP	Dense X Retrieval
Top-1						
Rec.	83.92	2.75	89.02	92.55	89.80	88.24
Pre.	95.11	2.75	89.02	92.55	89.80	88.24
IE	79.82	0.08	79.25	85.66	80.64	77.86
Top-3						
Rec.	87.84	5.10	96.08	96.08	96.08	95.29
Pre.	94.81	1.70	86.14	82.88	84.97	84.97
IE	83.28	0.09	82.76	79.63	81.64	80.97
Top-5						
Rec.	90.59	5.10	98.04	98.43	98.04	96.25
Pre.	95.14	1.02	82.70	69.57	77.65	82.75
IE	86.19	0.05	81.08	68.48	76.13	79.65

Table 8: **Detailed information of comparative experiments on NFCORPUS.**

Method	SAKI-RAG	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chuning -MSP	Dense X Retrieval
			Top-1			
Rec.	87.67	0.67	96.19	18.83	93.95	92.83
Pre.	97.51	0.67	96.19	18.83	93.95	92.83
IE	85.49	0.004	92.53	3.55	88.27	86.17
			Top-3			
Rec.	93.05	0.67	97.98	23.54	97.01	96.86
Pre.	97.55	0.22	92.68	19.28	88.27	87.29
IE	90.77	0.001	90.81	4.54	85.63	84.55
			Top-5			
Rec.	94.17	0.67	98.88	22.65	97.76	97.53
Pre.	97.46	0.13	90.49	16.64	82.69	84.48
IE	91.78	0.001	89.48	3.77	80.84	82.39

Table 9: **Detailed information of comparative experiments on SCI-DOCS.**

Method	SAKI-RAG	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chuning -MSP	Dense X Retrieval
			Top-1			
Rec.	32.62	0.57	30.16	40.43	32.30	2.71
Pre.	89.92	1.79	81.89	83.42	82.91	10.97
IE	29.33	0.01	24.70	33.73	26.78	0.30
			Top-3			
Rec.	49.94	0.88	49.81	64.48	52.83	3.97
Pre.	92.71	1.02	45.96	46.09	46.68	5.36
IE	46.30	0.01	22.89	29.72	24.66	0.21
			Top-5			
Rec.	61.90	1.07	57.62	73.68	61.78	4.16
Pre.	93.43	0.77	31.87	31.63	33.01	3.37
IE	57.83	0.01	18.36	23.30	20.39	0.14

Table 10: **Detailed information of comparative experiments on Dragonball-Hop.**

Method	SAKI-RAG	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chuning -MSP	Dense X Retrieval
			Top-1			
Rec.	9.31	0.38	8.55	12.68	8.09	0.80
Pre.	50.52	2.30	44.26	50.49	43.93	6.23
IE	4.70	0.01	3.78	6.40	3.55	0.05
			Top-3			
Rec.	22.54	0.51	19.63	29.53	17.65	1.31
Pre.	59.28	1.09	34.47	38.69	30.38	3.39
IE	13.36	0.01	6.77	11.43%	5.36	0.04
			Top-5			
Rec.	31.97	0.67	27.59	39.97	25.57	1.64
Pre.	61.32	0.85	28.82	31.21	26.82	2.56
IE	19.60	0.01	7.95	12.47	6.86	0.04

Table 11: **Detailed information of comparative experiments on Dragonball-Summary.**

Method	SAKI-RAG	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chuning -MSP	Dense X Retrieval
Top-1						
Rec.	18.63	0.45	17.21	23.80	17.79	1.56
Pre.	72.41	2.01	65.42	69.01	65.85	8.90
IE	13.49	0.01	11.26	16.42	11.71	0.14
Top-3						
Rec.	33.64	0.66	31.73	43.54	31.75	2.37
Pre.	78.16	1.05	40.94	42.85	39.55	4.50
IE	26.29	0.01	12.99	18.66	12.56	0.11
Top-5						
Rec.	44.07	0.83	39.63	53.48	40.08	2.65
Pre.	79.45	0.80	30.53	31.45	30.30	3.01
IE	35.01	0.01	12.10	16.82	12.14	0.08

Table 12: **Detailed information of comparative experiments on Dragonball-Non-Factual.**

Method	Naive	Naive+SAL	SAKI
Top-1			
Rec.	19.58	22.13	22.14
Pre.	69.36	69.68	74.84
IE	13.58	15.42	16.57
Top-3			
Rec.	33.22	36.89	36.96
Pre.	34.69	78.03	79.76
IE	11.52	28.79	29.48
Top-5			
Rec.	39.29	46.82	46.99
Pre.	24.56	79.59	80.72
IE	9.65	37.26	37.93

Table 13: **Detailed information of ablation studies on Dragonball.**

Method	Naive	Naive+SAL	SAKI
Top-1			
Rec.	86.87	86.87	86.33
Pre.	86.87	86.87	86.33
IE	75.46	75.46	79.85
Top-3			
Rec.	94.69	94.41	95.53
Pre.	35.75	89.11	95.18
IE	33.85	84.13	90.93
Top-5			
Rec.	92.25	96.65	95.54
Pre.	23.41	89.89	94.39
IE	21.60	86.88	90.18

Table 14: **Detailed information of ablation studies on SQUAD.**

Method	Naive	Naive+SAL	SAKI
Top-1			
Rec.	86.87	86.87	86.33
Pre.	86.87	86.87	86.33
IE	75.46	75.46	79.85
Top-3			
Rec.	94.69	94.41	95.53
Pre.	35.75	89.11	95.18
IE	33.85	84.13	90.93
Top-5			
Rec.	92.25	96.65	95.54
Pre.	23.41	89.89	94.39
IE	21.60	86.88	90.18

Table 15: **Detailed information of ablation studies on NFCORPUS.**

Method	Naive	Naive+SAL	SAKI
Top-1			
Rec.	91.93	93.50	87.67
Pre.	91.93	93.50	97.51
IE	84.51	87.42	85.49
Top-3			
Rec.	95.29	94.39	93.05
Pre.	87.97	93.72	97.55
IE	83.83	88.46	90.77
Top-5			
Rec.	95.96	94.84	94.17
Pre.	84.30	93.59	97.46
IE	80.89	88.76	91.78

Table 16: **Detailed information of ablation studies on SCI-DOCS.**

Dragonball	SAKI-RAG -Qwen -max	SAKI-RAG -Qwen -turbo	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chunking -MSP	Dense X Retrieval
Top-1							
Recall	22.14	22.19	0.52	17.21	26.75	20.71	1.81
Precision	74.73	72.65	1.83	65.42	68.67	64.50	7.83
F1	0.34	0.34	0.01	0.27	0.39	0.31	0.03
EIR	62.94	62.89	71.30	75.60	51.18	63.97	99.20
MRR	0.70	0.70	0.02	0.65	0.69	0.64	0.08
Time Ave	0.83	0.49	0.04	0.03	0.03	0.03	0.03
Top-3							
Recall	36.99	36.94	0.75	31.73	45.93	34.54	2.66
Precision	79.66	79.04	0.95	40.94	38.08	35.16	3.83
F1	0.51	0.50	0.01	0.36	0.42	0.35	0.03
EIR	16.85	16.88	32.04	35.41	24.33	29.30	42.00
MRR	0.81	0.81	0.01	0.59	0.75	0.71	0.09
Time Ave	6.04	4.04	0.04	0.04	0.03	0.03	0.03
Top-5							
Recall	46.94	46.89	0.96	39.63	55.53	42.61	3.01
Precision	80.63	80.27	0.73	30.53	27.20	26.02	2.60
F1	0.59	0.59	0.01	0.34	0.37	0.32	0.03
EIR	11.27	11.27	24.09	24.46	16.78	20.06	26.37
MRR	0.84	0.83	0.01	0.53	0.76	0.72	0.10
Time Ave	16.57	11.03	0.04	0.06	0.03	0.03	0.03

Table 17: **Comparative experiments on Dragonball with more metrics.**

NFCORPUS	SAKI-RAG -Qwen -max	SAKI-RAG -Qwen -turbo	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chunking -MSP	Dense X Retrieval
Top-1							
Recall	85.49	87.84	2.75	89.02	92.55	89.80	88.24
Precision	94.78	94.12	2.75	89.02	92.55	89.80	88.24
F1	0.90	0.91	0.03	0.89	0.93	0.90	0.88
MRR	0.85	0.88	0.03	0.89	0.93	0.90	0.88
Time Ave	0.67	0.48	0.03	0.04	0.03	0.03	0.03
Top-3							
Recall	90.59	92.94	5.10	96.08	95.69	96.08	95.29
Precision	95.20	94.58	1.02	86.14	82.75	84.97	84.97
F1	0.93	0.94	0.02	0.91	0.89	0.90	0.90
MRR	0.90	0.92	0.04	0.91	0.94	0.93	0.92
Time Ave	5.92	3.87	0.03	0.07	0.03	0.03	0.03
Top-5							
Recall	90.59	92.94	5.10	98.04	98.43	97.65	97.25
Precision	94.93	94.70	1.02	82.70	69.49	77.41	82.75
F1	0.93	0.94	0.02	0.90	0.81	0.86	0.89
MRR	0.90	0.93	0.03	0.90	0.94	0.93	0.92
Time Ave	18.62	11.58	0.03	0.10	0.03	0.03	0.03

Table 18: **Comparative experiments on NFCORPUS with more metrics.**

SQUAD	SAKI-RAG -Qwen -max	SAKI-RAG -Qwen -turbo	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chunking -MSP	Dense X Retrieval
Top-1							
Recall	86.87	86.59	16.48	/	78.21	79.33	56.98
Precision	91.20	90.64	16.48	/	78.21	79.33	56.98
F1	0.89	0.89	0.16	/	0.78	0.79	0.57
MRR	0.87	0.87	0.16	/	0.78	0.79	0.57
Time Ave	0.66	0.51	0.02	/	0.03	0.03	0.03
Top-3							
Recall	85.20	/	25.70	/	86.87	87.99	71.79
Precision	91.32	/	10.61	/	32.40	32.77	28.12
F1	0.88	/	0.15	/	0.47	0.48	0.40
MRR	0.85	/	0.16	/	0.84	0.84	0.64
Time Ave	0.63	/	0.02	/	0.03	0.03	0.03
Top-5							
Recall	/	/	28.49	/	87.71	89.11	75.42
Precision	/	/	7.49	/	20.39	20.67	19.11
F1	/	/	0.12	/	0.33	0.34	0.30
MRR	/	/	0.13	/	0.84	0.84	0.65
Time Ave	/	/	0.03	/	0.03	0.03	0.03

Table 19: **Comparative experiments on SQUAD with more metrics.**

SCI-DOCS	SAKI-RAG -Qwen -max	SAKI-RAG -Qwen -turbo	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chunking -MSP	Dense X Retrieval
Top-1							
Recall	89.69	91.70	0.37	96.19	19.73	93.65	92.83
Precision	97.56	91.70	0.37	96.19	19.73	93.65	92.83
F1	0.93	0.94	0.01	0.96	0.20	0.94	0.93
MRR	0.90	0.92	0.01	0.96	0.20	0.94	0.93
Time Ave	0.69	0.52	0.09	0.08	0.03	0.03	0.03
Top-3							
Recall	93.72	94.39	0.67	97.98	22.65	97.09	96.86
Precision	97.57	96.75	0.22	92.68	18.61	88.34	87.29
F1	0.96	0.96	0.003	0.95	0.20	0.93	0.92
MRR	0.93	0.94	0.01	0.95	0.21	0.95	0.95
Time Ave	5.80	4.00	0.08	0.13	0.03	0.03	0.03
Top-5							
Recall	93.72	94.84	0.67	98.88	23.32	97.76	97.76
Precision	97.27	96.61	0.13	90.49	17.35	82.78	84.57
F1	0.95	0.96	0.002	0.95	0.20	0.90	0.91
MRR	0.93	0.94	0.01	0.95	0.21	0.96	0.95
Time Ave	16.48	12.08	0.08	0.22	0.03	0.03	0.03

Table 20: **Comparative experiments on SCI-DOCS with more metrics.**

HotpotQA	SAKI-RAG -Qwen -max	SAKI-RAG -Qwen -turbo	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chunking -MSP	Dense X Retrieval
Top-1							
Recall	86.71	89.24	3.80	/	6.33	93.04	93.04
Precision	98.56	68.60	3.80	/	6.33	93.04	93.04
F1	0.92	0.94	0.04	/	0.06	0.93	0.93
MRR	0.87	0.89	0.04	/	0.06	0.93	0.93
Time Ave	0.68	0.43	0.05	/	0.03	0.03	0.03
Top-3							
Recall	94.94	93.04	3.80	/	5.70	93.67	96.84
Precision	99.11	98.19	2.11	/	3.59	78.90	89.66
F1	0.97	0.96	0.03	/	0.04	0.86	0.93
MRR	0.95	0.93	0.04	/	0.05	0.93	0.95
Time Ave	5.58	3.86	0.05	/	0.03	0.03	0.03
Top-5							
Recall	94.94	94.94	3.80	/	12.66	94.30	96.84
Precision	98.79	97.86	1.90	/	6.33	66.84	87.59
F1	0.97	0.96	0.03	/	0.08	0.78	0.92
MRR	0.99	0.95	0.04	/	0.33	0.94	0.95
Time Ave	20.44	10.80	0.05	/	0.03	0.03	0.03

Table 21: **Comparative experiments on HotpotQA with more metrics.**

TriviaQA	SAKI-RAG -Qwen -max	SAKI-RAG -Qwen -turbo	Late-Chunking	RAPTOR	Meta-Chunking -PPL	Meta-Chunking -MSP	Dense X Retrieval
Top-1							
Recall	65.25	64.09	46.72	/	3.47	57.92	35.91
Precision	76.47	75.11	46.72	/	3.47	57.92	35.91
F1	0.70	0.69	0.47	/	0.03	0.58	0.36
MRR	0.65	0.64	0.47	/	0.03	0.58	0.05
Time Ave	0.70	0.51	0.05	/	0.03	0.03	0.03
Top-3							
Recall	84.56	84.94	71.43	/	6.18	74.90	57.53
Precision	84.91	83.50	34.36	/	2.70	36.98	30.24
F1	0.85	0.84	0.46	/	0.04	0.49	0.40
MRR	0.81	0.81	0.51	/	0.01	0.66	0.07
Time Ave	6.31	4.71	0.04	/	0.03	0.03	0.03
Top-5							
Recall	88.80	89.96	80.79	/	5.79	77.61	66.80
Precision	85.08	83.87	30.12	/	1.85	26.41	26.53
F1	0.87	0.87	0.44	/	0.03	0.39	0.38
MRR	0.84	0.85	0.47	/	0.01	0.67	0.08
Time Ave	19.39	11.15	0.04	/	0.03	0.03	0.03

Table 22: **Comparative experiments on TriviaQA with more metrics.**

Dragonball Generation Quality	SAKI-RAG	Light-RAG
Relevant	70.99	72.97
Irrelevant	27.01	22.99
Wrong	2.01	4.04

Table 23: **Comparative experiments on Generation with more metrics.**

Dragonball	Meta-Chunking -MSP	Meta-Chunking -PPL	RAPTOR
Recall@3 Mean Difference	0.045	-0.053	0.047
Recall@3 two-tailed p	0.000033★	0.000033★	0.000033★
Precision@3 Mean Difference	0.43	0.401	0.423
Precision@3 two-tailed p	0.000033★	0.000033★	0.000033★
EIR@3 Mean Difference	-0.087	-0.066	-0.1
EIR@3 two-tailed p	0.000033★	0.000033★	0.000033★
F1@3 Mean Difference	0.25	0.197	0.248
F1@3 two-tailed p	0.000033★	0.000033★	0.000033★
MRR@3 Mean Difference	0.106	0.061	0.277
MRR@3 two-tailed p	0.000033★	0.000033★	0.000033★

Table 24: **Statistical validation experiments on Dragonball.**

Dragonball	Meta-Chunking -MSP	Meta-Chunking -PPL	RAPTOR	DXR
Recall@3 Mean Difference	-0.035	-0.035	-0.035	-0.027
Recall@3 two-tailed p	0.08	0.00007★	0.05	0.208
Precision@3 Mean Difference	0.061	0.084	0.051	0.063
Precision@3 two-tailed p	0.001★	0.00007★	0.006★	0.002
F1@3 Mean Difference	0.031	0.046	0.022	0.033
F1@3 two-tailed p	0.081	0.006★	0.183	0.084
MRR@3 Mean Difference	-0.001	-0.024	0.018	0.004
MRR@3 two-tailed p	0.665	0.224	0.310	0.859

Table 25: **Statistical validation experiments on NFCORPUS.**

Dragonball	Meta-Chunking -MSP	Meta-Chunking -PPL	RAPTOR	DXR
Recall@3 Mean Difference	-0.031	0.709	-0.04	-0.029
Recall@3 two-tailed p	0.004★	0.00003★	0.0004★	0.019★
Precision@3 Mean Difference	0.05	0.744	0.006	0.06
Precision@3 two-tailed p	0.0003★	0.00003★	0.57	0.00003★
F1@3 Mean Difference	0.025	0.735	0.011	0.03
F1@3 two-tailed p	0.045★	0.00003★	0.329	0.018★
MRR@3 Mean Difference	-0.019	0.714	-0.017	-0.013
MRR@3 two-tailed p	0.116	0.00003★	0.169	0.32

Table 26: **Statistical validation experiments on SCI-DOCS.**

Dragonball	Meta-Chunking -MSP	Meta-Chunking -PPL	DXR
Recall@3 Mean Difference	-0.006	0.816	-0.029
Recall@3 two-tailed p	1	0.00003★	0.019★
Precision@3 Mean Difference	0.135	0.85	0.06
Precision@3 two-tailed p	0.00003★	0.00003★	0.00003★
F1@3 Mean Difference	0.093	0.837	0.03
F1@3 two-tailed p	0.00007★	0.00003★	0.018★
MRR@3 Mean Difference	-0.003	0.824	-0.013
MRR@3 two-tailed p	1	0.00003★	0.32

Table 27: **Statistical validation experiments on HotpotQA.**

Dragonball	Meta-Chunking -MSP	Meta-Chunking -PPL	DXR
Recall@3 Mean Difference	0.085	0.788	0.263
Recall@3 two-tailed p	0.002★	0.00003★	0.00003★
Precision@3 Mean Difference	0.393	0.743	0.459
Precision@3 two-tailed p	0.00003★	0.00003★	0.00003★
F1@3 Mean Difference	0.311	0.757	0.403
F1@3 two-tailed p	0.00003★	0.00003★	0.00003★
MRR@3 Mean Difference	0.659	0.792	0.726
MRR@3 two-tailed p	0.00003★	0.00003★	0.00003★

Table 28: **Statistical validation experiments on TriviaQA.**

Dragonball	
ROUGE-L (two-tailed p)	4.51e-13★
BLEU-1 (two-tailed p)	5.83e-21★
BLEU-2 (two-tailed p)	4.62e-23★
BLEU-3 (two-tailed p)	1.40e-05★
BLEU-4 (two-tailed p)	0.007★

Table 29: **Statistical validation experiments on generation quality.**

{'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 5: 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 6: 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 7: Detailed informations of SCI-DOCS dataset.

{"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 8: Detailed informations of Dragonball dataset.