Concept-Based RAG Models: A High-Accuracy Fact Retrieval Approach

Cheng-Yu Lin
Department of Computer Science
National Taiwan University, Taiwan
d12922010@csie.ntu.edu.tw

Jyh-Shing Roger Jang
Department of Computer Science
National Taiwan University, Taiwan
jang@csie.ntu.edu.tw

Abstract

This study introduces a concept-based methodology to optimize Retrieval-Augmented Generation (RAG) tasks by assessing dataset certainty using entropy-based metrics and concept extraction techniques. Unlike traditional methods focused on reducing LLM hallucinations or modifying data structures, this approach evaluates inherent knowledge uncertainty from an LLM perspective. By pre-processing documents with LLMs, the concept-based method significantly enhances precision in tasks demanding high accuracy, such as legal, finance, or formal document responses.

1 Introduction

Retrieval-Augmented Generation (RAG) is an advanced framework that combines generative AI models with external retrieval capabilities to provide answers with higher accuracy and contextual relevance.

This paper introduces several common types of RAG models and analyzes their core features, application scenarios, and technical architecture. These types include Standard RAG (Lewis et al., 2021), Corrective RAG (Yan et al., 2024), Graph RAG (Edge et al., 2024), Agentic RAG (Ravuru et al., 2024), and Dynamic Hierarchical RAG (Wang et al., 2024), each showcasing powerful retrieval and generation capabilities across different application domains, catering to diverse information needs. However, these RAG models are constrained by the requirement that the dataset must have consistent and meaningful content throughout (low entropy); otherwise, these RAG models cannot ensure that the referenced information is derived from the correct documents.

1.1 Entropy

In information theory, Shannon entropy (Shannon, 1948) is defined as the uncertainty associated with a random variable.

In the context of large language models (LLMs), the generated text can be seen as comprising multiple concepts $C_1, C_2, \ldots, C_N$, whose probabilities are influenced by the context S. The entropy of a document D can be expressed as:

$$H(D) = -\sum_{i=1}^{N} p(C_i \mid S) \log p(C_i \mid S),$$

where $p(C_i \mid S)$ is the conditional probability of concept C_i under a specific context S. This extension reveals that entropy reflects not only the diversity of the concepts but also how context influences the content generated by the model.

As Shannon stated, "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." (Shannon, 1948) Based on this principle, the entropy of a document can be simplified to the idea that each document D contains multiple potential concepts $C_1, \ldots, C_N$, whose entropy is computed depending on the specific usage context S.

1.2 LLM and RAG

Transformer-based models (Vaswani et al., 2023) leverage scalable self-attention mechanisms to effectively encode complex linguistic information, achieving significant success across various Natural Language Processing (NLP) tasks without requiring extensive labeled data. In particular, the GPT series (Radford et al., 2019), as Decoder-based architectures, generate probabilistic textual outputs, enabling novel possibilities for Retrieval-Augmented Generation (RAG) frameworks powered by Large Language Models (LLMs). The following equations outline the inference and generation process:

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 96–100
January 19–20, 2025.

1.2.1 RAG Process

The process can be described as:

$$\text{LLM} : (C, Q) \xrightarrow{\text{Attention}} \text{Conceptual Alignment} \xrightarrow{\text{Generation}} A$$

- **Attention**: Achieves alignment between context and question concepts using self-attention.
- **Conceptual Alignment**: Serves as a channel for aligning concepts between context and question. This step can also be replaced by term-based models, such as BM25 (Robertson and Zaragoza, 2009), which offer explicit term matching. Since LLMs may inject implicit or irrelevant concepts, alternative approaches might provide more reliable alignments.
- **Generation**: Produces the final answer based on aligned concepts.

2 Related Work

RAG technologies aim to improve retrieval and generation through various methods in technical applications, yet each faces unique challenges. Standard RAG improves retrieval speed and precision, but struggles with accuracy and coordination with generative models. Corrective RAG enables continuous learning, but struggles with balancing speed and stability. Graph RAG leverages knowledge graphs for logical reasoning, but is hindered by design complexity and scenario diversity. Agentic RAG focuses on integrating multiple knowledge bases but faces difficulty creating adaptable reasoning frameworks. DML RAG excels in dynamic adaptation but struggles with maintaining accuracy and interpretability.

A shared challenge across these approaches is managing the complexity of input data. Models often fail to effectively extract key information from large, diverse datasets, reducing the reliability of generated outputs. Improving input data processing and improving collaboration between retrieval and generation is critical to advancing RAG technologies.

3 Methodology

3.1 Background

As the number of concepts n in the context grows, the language model must manage more interdependencies, increasing uncertainty reflected in conditional entropy. The conditional entropy $H(A \mid Q, C, \text{LLM})$ is defined as:

$$H(A \mid Q, C, \text{LLM}) = -\sum_{a \in \mathcal{A}} P(a \mid Q, C, \text{LLM}) \log P(a \mid Q, C, \text{LLM}),$$

where $C = \{C_1, C_2, \ldots, C_n\}$ is the set of concepts, and $P(a \mid Q, C, \text{LLM})$ is the probability of generating answer a given query Q and context C.

3.2 Method

To evaluate the informational richness of an article, this study employs a concept-based metric. By utilizing large language models (LLMs) in combination with carefully designed prompts, individual conceptual segments are extracted from the text. The methodology is outlined as follows:

1. **Conceptual Segment Extraction**: Prompts generated by the LLM are used to extract fragments from the text, each representing a distinct individual concept.
2. **Entropy Calculation**: The number of extracted segments is used as the basis for computing the entropy of the article.

For simplicity, we assume that all segment probabilities are equal. While these probabilities may vary due to contextual alignment, such variations are beyond the scope of this discussion. Consequently, the entropy function primarily depends on the number of conceptual segments contained within the article.

3.3 Workflow Overview

The proposed workflow begins by inputting an article into the LLM, using prompts specifically designed to ensure that each output segment represents a single distinct concept. This segmentation process breaks down the article into smaller, more concise fragments. These fragments are then incorporated into the Retrieval-Augmented Generation (RAG) pipeline for subsequent processing.

For contextual alignment, this study adopts the BM25 algorithm to evaluate and rank the extracted segments. The complete processing workflow is visualized in Figure 1.

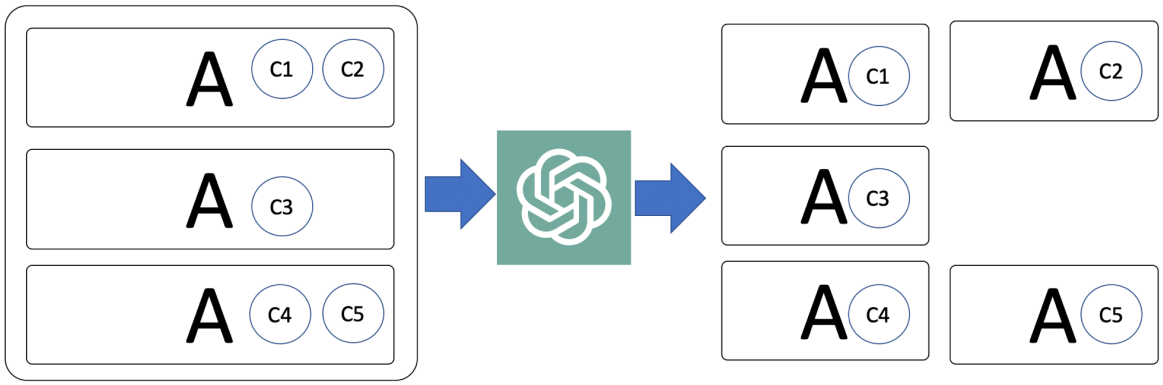

Figure 1: An example illustrating the process of segmenting an article into smaller fragments and integrating them into the Retrieval-Augmented Generation (RAG) workflow.

4 Experiment

4.1 Datasets

This study addresses financial legal documents requiring jurisdiction-specific localization, using a dataset from a public competition in Taiwan(TBrain) featuring 1,038 corporate financial reports and 300 finance-related questions. Approximately 500,000 words are extracted using 'pdfplumber' and GPT 4o-mini, with documents evaluated through full-text retrieval, restricted-scope retrieval, and a concept-based approach. Conceptual fragments, extracted via tailored prompts (Figure 2), enhance alignment and relevance in financial question-answering tasks.

4.2 Evaluation Metrics

The competition evaluates retrieval performance using the Precision@1 score, which measures the accuracy of the top-ranked retrieved document for each query. The formula for Precision@1 is defined as follows:

$$\text{Precision@1} = \frac{\text{Top 1 Documents}}{\text{Ground Truth Documents}}$$

For the preliminary evaluation, the Average Precision@1 is used as the overall performance metric. This metric calculates the mean Precision@1 across all queries, rounded to seven decimal places. An example is provided below for clarification: Precision@1 for each query is calculated as follows:

- Precision@1 for Q1: $\frac{1}{1} = 1.0$
- Precision@1 for Q2: $\frac{0}{1} = 0.0$

Query	Predicted Result	Ground Truth
Q1	D1	D1
Q2	D2	D3
Q3	D3	D3

Table 1: Example illustrating Precision@1 calculation.

- Precision@1 for Q3: $\frac{1}{1} = 1.0$

The Average Precision@1 is then computed as:

$$\text{Average Precision@1} = \frac{(1.0 + 0.0 + 1.0)}{3} = 0.67$$

This metric provides a straightforward and reliable measure for assessing retrieval accuracy in the competition and the experiment.

4.3 Comparison of Retrieval Methods

The experimental results clearly demonstrate that the concept-based BM25 significantly outperforms the traditional BM25 in financial retrieval tasks. As shown in Table 2, the concept-based BM25 achieves a 33% improvement in Precision for partial retrieval (0.64 vs. 0.48) and a remarkable improvement for full retrieval, where Precision increases from 0.08 to 0.30.

As shown in Table 3, the analysis of Entropy further illustrates the advantages of the concept-based BM25 method. For partial retrieval scenarios, the Entropy value decreases to 0.13, reflecting higher clarity of document information within a smaller search scope and significantly improved semantic consistency in the retrieval results. This aligns with intuitive understanding, where a smaller scope leads to more concentrated and clearer information. In contrast, for full retrieval scenarios, the Entropy value increases to 0.33, demonstrating that

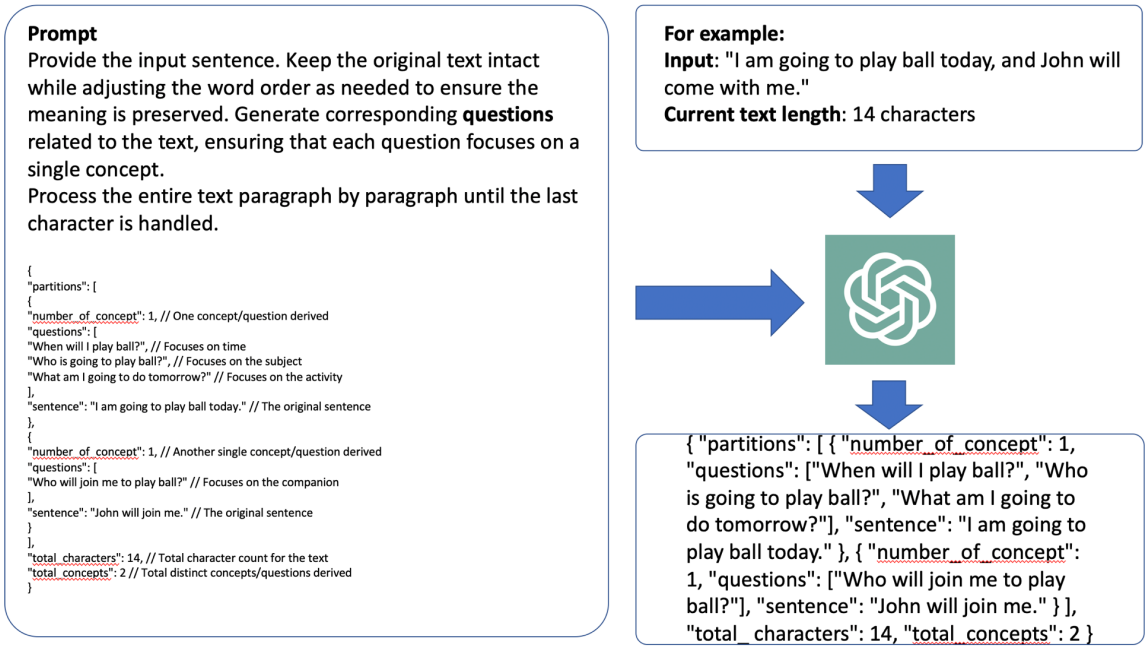

Figure 2: Illustration of the process for extracting conceptual fragments from documents and integrating them into the Retrieval-Augmented Generation (RAG) pipeline. This workflow demonstrates how financial documents are segmented and processed for improved alignment and retrieval accuracy. For detailed prompts and experimental setup, please refer to Appendix A.

Search Strategy	Traditional BM25 Precision	Concept-based BM25 Precision
Partial	0.48	0.64
Full	0.08	0.30

Table 2: Comparison of Precision between traditional BM25 and concept-based BM25.

Search Strategy	Entropy
Partial	0.13
Full	0.33

Table 3: Comparison of entropy between two types of strategies.

the concept-based BM25 method effectively handles complex document structures and accurately identifies key information in large-scale corpora.

These findings emphasize the concept-based BM25 method's superior sensitivity to semantic features and its enhanced ability to understand and utilize semantic hierarchies. Such improvements are particularly critical for financial applications like question-answering systems and information retrieval tasks. The results confirm that adopting the concept-based BM25 method effectively enhances retrieval performance in the financial domain.

5 Conclusions and Future work

The results confirm that the proposed concept-based BM25 method significantly enhances the precision of term-based models (such as BM25) in semantic matching tasks. This demonstrates the effectiveness of integrating conceptual segmentation as a pre-processing step to address semantic alignment challenges.

In future work, we plan to incorporate document entropy as an additional evaluation metric. This will enable more sophisticated selection and utilization of datasets for vector-based or graph-based retrieval methods, further improving the accuracy of selecting relevant documents for generation tasks.

Notably, the proposed method operates as a pre-processing step and does not occupy inference time, making it highly practical and efficient for integration into existing retrieval-augmented generation workflows.

Acknowledgments

We would like to express our gratitude to the organizers of the competition for providing valuable information and resources that greatly supported this study. Their assistance was instrumental in enabling a thorough evaluation of the proposed methods.

To facilitate further research and ensure reproducibility, we plan to release the relevant code and datasets. The associated code can be accessed via the following link: https://github.com/ntuaha/Concept_Based_RAG.

References

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.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Chidaksh Ravuru, Sagar Srinivas Sakhinana, and Venkataramana Runkana. 2024. Agentic retrieval-augmented generation for time series analysis. *Preprint*, arXiv:2408.14484.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

TBrain. 2023 finance question answering competition. https://tbrain.trendmicro.com.tw/Competitions/Details/37. Accessed: 2024-11-24.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Xinyu Wang, Yanzheng Xiang, Lin Gui, and Yulan He. 2024. Garlic: Llm-guided dynamic progress control with hierarchical weighted graph for long document qa. *Preprint*, arXiv:2410.04790.

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

A Prompt

This appendix provides the detailed prompts used in the original experiments. The prompts are designed to ensure effective segmentation and accurate retrieval during the implementation of the RAG pipeline.

To enable more precise segmentation of documents in Traditional Chinese, we utilize the following prompt for segmentation in Figure 3. This prompt divides the documents into individual "Concepts," which are subsequently used for document alignment and retrieval tasks. (Figures in the main text illustrate the process in English.)

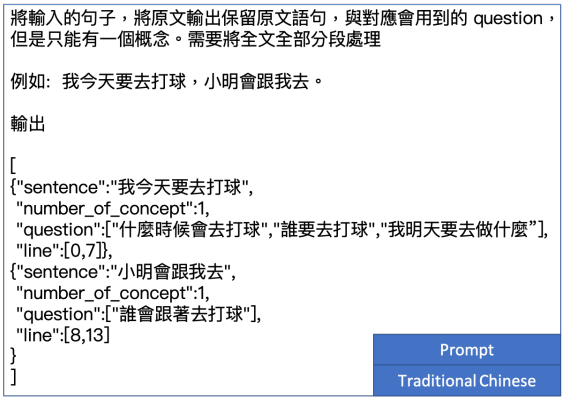

Figure 3: Prompt in Traditional Chinese for segmenting financial documents into conceptual fragments.