

Enhancing Thai Annual Report Queries with Retrieval-Augmented Generation

Burin Deesiri ¹
Noptanit Chotisarn ^{2,*}

*Corresponding author

¹ Management Information System Department, Thammasat Business School, Thammasat University, Bangkok, Thailand, burin-dee65@tbs.tu.ac.th

² Management Information System Department, Thammasat Business School, Thammasat University, Bangkok, Thailand, noptanit@tbs.tu.ac.th

ABSTRACT

Analyzing annual reports of Thai public companies poses significant challenges due to the complexity and volume of unstructured textual data. To address this, we propose a Retrieval-Augmented Generation (RAG) framework that integrates dense retrieval techniques with large language models (LLMs) to facilitate natural language querying over financial documents. The framework utilizes the GTE-Large embedding model to semantically encode and retrieve relevant document segments from a vector store, which are then used by the Llama 3-8B model to generate informed, context-specific responses. A qualitative comparison between the RAG system and a naive LLM setup reveals that the RAG approach significantly reduces hallucinations and produces outputs that are more faithful to the source documents. The results underscore the importance of grounding language model outputs in external knowledge when dealing with domain-specific queries. This lightweight and modular framework demonstrates practical potential for scalable applications in financial analytics and similar domains requiring accurate information extraction. Future research will extend this work by incorporating advanced optimization techniques, deploying quantitative evaluation metrics to benchmark performance, and exploring real-world deployment strategies for interactive financial question-answering systems.

Keywords: Natural language processing, retrieval-augmented generation, annual reports, large language models, financial data analysis.

INTRODUCTION

The growing volume of unstructured data in the business world presents major challenges for effective analysis. Among these, annual reports from public companies remain a vital yet underutilized resource. These reports contain critical financial and strategic information, but their sheer length and complexity make them difficult to interpret, especially for time-constrained stakeholders like investors. Marston and Shrides (1991) emphasize the annual report as a key disclosure tool for shaping public perception and investment decisions.

In Thailand, public companies listed on the Stock Exchange of Thailand (SET) are required to publish their annual reports using the standardized “56-1 One Report” format (Juntrakool, 2023). Despite this structure, reports often exceed 200 pages and are commonly released in PDF format, making them difficult to search or analyze systematically. Online demand for these reports remains high, with more than 5,000 monthly searches for “รายงานประจำปี” (annual report) in 2023, according to Google Keyword Planner (Google, 2023).

Recent developments in Natural Language Processing (NLP) and Large Language Models (LLMs) have shown potential to transform how users interact with unstructured documents. Tools like ChatGPT, based on transformer architectures (Vaswani *et al.*, 2017), can answer complex questions and summarize information in human-like language. However, such models also face limitations: their responses may be based on outdated or incomplete data, they cannot cite sources, and they may generate “hallucinations” — confident but incorrect statements (Lewis *et al.*, 2020).

To address these shortcomings, Retrieval-Augmented Generation (RAG) frameworks have emerged. RAG combines the generative power of LLMs with real-time document retrieval, offering both fluency and factual grounding (Cui *et al.* (2023). This approach enables AI systems to answer questions based on up-to-date, traceable information rather than static, internalized knowledge alone.

This study applies a RAG-based approach tailored to the Thai financial reporting context. By connecting language models to a curated dataset of 243 annual reports from SET50 companies, our system enables users to ask natural-language questions and receive fact-grounded, verifiable answers. The goal is to enhance accessibility and insight extraction for financial stakeholders while reducing hallucinations and improving trust in AI-generated content.

To evaluate the effectiveness of this approach, we conducted a qualitative comparison between the RAG pipeline and a naive LLM setup using authentic queries from real Thai annual reports. The findings, presented later in the Results and Discussion

section, illustrate clear improvements in factuality and domain alignment. While the study prioritizes interpretability over strict numerical benchmarks, it lays a foundation for more rigorous future evaluation.

BACKGROUND ON ANNUAL REPORTS AND CORPORATE DISCLOSURE

Annual reports are essential documents through which public companies communicate financial performance, strategic direction, and risk disclosures to a wide range of stakeholders. As emphasized by Marston and Shrivies (1991), annual reports serve as a company's primary disclosure vehicle, shaping investor perceptions and supporting corporate accountability. These reports are particularly valuable to investors, who rely on them to assess business health and make informed decisions.

In Thailand, public companies listed on the Stock Exchange of Thailand (SET) are required to submit annual reports using the standardized "56-1 One Report" format (Juntrakool, 2023). This structure ensures a degree of consistency across firms and includes detailed disclosures on financial results, risk factors, governance, and strategic outlook. Despite these benefits, annual reports are often lengthy—ranging from 200 to 400 pages—and presented in PDF format, limiting accessibility and ease of use.

While standardized formats offer a useful baseline, companies retain discretion over section ordering and labeling, which complicates automated comparisons across firms and years. In addition, these reports include a mix of text, tables, charts, and images, creating challenges for both human readers and traditional data extraction tools (Stanton & Stanton, 2002). Extracting actionable insights from these complex documents remains a significant bottleneck for analysts and investors.

This challenge is magnified when attempting to analyze reports across multiple years or companies. As noted by El-Haj *et al.* (2019), even advanced computational systems face difficulties when dealing with PDF-encoded financial documents due to structural inconsistencies and content heterogeneity. In the case of Thailand, growing interest in financial transparency is reflected in rising search volumes for "annual reports" (รายงานประจำปี) among public users (Google, 2023).

Recent advances in Natural Language Processing (NLP) offer a promising solution to these challenges. NLP enables automated understanding of unstructured textual data, and its adoption in financial reporting analysis has increased steadily in recent years. In this study, we propose a Retrieval-Augmented Generation (RAG) approach that leverages NLP to extract and summarize content from annual reports issued by Thai public companies.

NATURAL LANGUAGE PROCESSING AND LARGE LANGUAGE MODELS

Natural Language Processing (NLP) is a core area of artificial intelligence that enables machines to understand and generate human language. In its early stages, NLP relied on statistical techniques like Bag-of-Words (BoW) and TF-IDF to analyze text data (Ramos *et al.*, 2003; Zhang *et al.*, 2010). However, these methods could not capture contextual meaning or semantic relationships between words. As the field matured, neural network-based embeddings like Word2Vec and GloVe introduced the concept of distributed word representations (Mikolov *et al.*, 2013; Pennington *et al.*, 2014), allowing words with similar meanings to be positioned close together in vector space.

A major leap in NLP came with the introduction of transformer-based models (Vaswani *et al.*, 2017), which allowed for more accurate context-aware understanding of language. The Transformer architecture underpins many recent advances in the field, including BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer) models. BERT improved text comprehension by analyzing sentences bidirectionally (Kenton & Toutanova, 2019), while GPT-3, developed by OpenAI, demonstrated few-shot learning capabilities and text generation at scale with 175 billion parameters (Brown, 2020).

These large-scale models, commonly referred to as Large Language Models (LLMs), are now capable of performing complex tasks such as summarization, translation, and question answering. However, they still face key limitations: their knowledge is fixed at the time of training, and they can sometimes generate incorrect or unverifiable information—a phenomenon known as hallucination (Lewis *et al.*, 2020).

To mitigate these issues, Retrieval-Augmented Generation (RAG) has been proposed as an enhanced framework that integrates external document retrieval into the generative process (Zhao *et al.*, 2023). This allows the model to ground its answers in up-to-date and verifiable sources, improving factual accuracy. In the context of financial reporting, especially for domain-specific and evolving content such as Thai corporate annual reports, RAG offers a practical solution for improving the reliability and interpretability of generated responses.

By leveraging recent advances in NLP and LLMs, this research aims to demonstrate the effectiveness of RAG for real-world financial text analysis. The focus is on enabling end-users—particularly investors and analysts—to extract meaningful insights from Thai public company reports with reduced cognitive effort and increased confidence in the accuracy of the information.

RETRIEVAL-AUGMENTED GENERATION (RAG)

Large Language Models (LLMs) have brought impressive breakthroughs in natural language processing but still face persistent challenges. Their reliance on static knowledge—embedded during training—limits their ability to provide up-to-date, verifiable responses. They can also generate "hallucinated" content that appears confident but lacks factual grounding. To address these

issues, the Retrieval-Augmented Generation (RAG) framework was introduced as a hybrid solution that enhances LLMs by integrating real-time document retrieval.

RAG improves reliability and transparency by retrieving relevant context from external sources, such as document repositories or databases, before generating a response. This approach allows the model to reference dynamic, traceable information rather than relying solely on internalized knowledge (Lewis *et al.*, 2020). For example, in the financial domain, where regulations and data change frequently, RAG can deliver responses grounded in the most recent disclosures or filings—an advantage over traditional LLMs.

In practice, RAG has been successfully applied across several domains, including law, healthcare, and finance. ChatLaw, a legal AI system developed in China, uses RAG to improve the interpretation of legal texts and enhance trust in machine-generated advice (Cui *et al.*, 2023). Similarly, in the financial sector, researchers have used RAG to improve sentiment analysis and question-answering over corporate disclosures and market news (B. Zhang *et al.*, 2023).

The RAG framework typically consists of two components:

- A retriever, which searches a large corpus of documents using either keyword-based (sparse) or vector-based (dense) methods to find the most relevant passages;
- A generator, usually a pre-trained LLM such as BART or LLaMA2, which takes the retrieved content along with the user's query to produce an answer (Touvron *et al.*, 2023).

This study adopts RAG for a specific use case: enabling investors and analysts to query Thai public company annual reports in natural language. The goal is to offer more transparent, grounded, and contextually precise insights from lengthy and complex financial documents—thereby improving both accessibility and decision-making quality in corporate finance.

METHODOLOGY

This study adopts a Retrieval-Augmented Generation (RAG) approach to develop a question-answering system tailored to Thai public company annual reports. The methodology focuses on transforming unstructured financial data into a searchable format that supports natural language interaction. By integrating data retrieval and generation processes, the framework provides accurate, context-rich answers based on real company disclosures. The following subsections outline each component of the system, starting with a high-level view of the proposed framework.

Proposed Retrieval-Augmented Generation (RAG) Framework

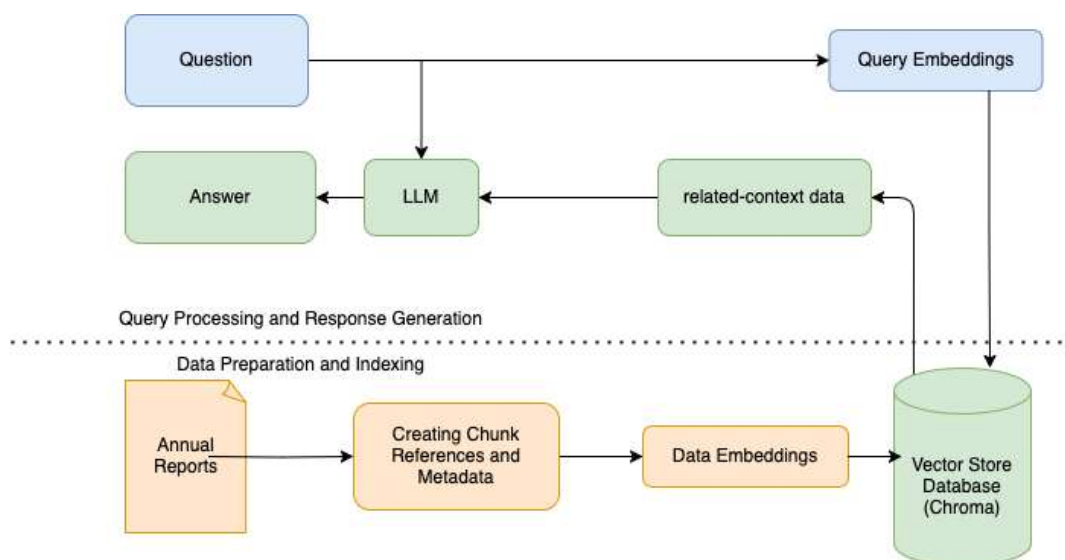


Figure 1: Proposed Retrieval-Augmented Generation (RAG) Framework.

The core framework consists of four main components that work together to support user queries over annual report content. As shown in Figure 1, the process begins with **data collection and preprocessing**, where financial reports are parsed and structured into meaningful text chunks. These chunks are then converted into vector representations using an **embedding model**, allowing the system to measure semantic similarity between user questions and the underlying content. The embeddings are stored in a **vector store database**, which acts as a fast, searchable index of relevant content. Finally, when a question is asked, a **large language model (LLM)** retrieves the most relevant information and generates a coherent, human-readable response. This architecture is designed to overcome common challenges in LLMs such as outdated knowledge and hallucination, particularly when applied to dynamic financial documents.

Data Collection and Preprocessing

The first step in the proposed RAG framework involves collecting and preparing textual data from annual reports published by Thai public companies. These reports, sourced from the Securities and Exchange Commission (SEC) and corporate websites, are typically provided in PDF format and contain extensive financial, strategic, and governance information.

To convert this unstructured content into a machine-readable format, the reports are processed through a structured pipeline. The pipeline begins by extracting plain text from the PDF files using automated tools. After extraction, the text is divided into manageable chunks of around 512 tokens each. This chunking method preserves contextual meaning while ensuring that the data can be efficiently embedded and retrieved later.

Each chunk is enriched with metadata such as the document name, year, and page number to support traceability during the retrieval process. This metadata allows the system to reference the source of each retrieved segment and improves transparency when presenting answers to users.

The dataset spans five years (2019–2023) and includes 242 reports totaling more than 29 million words. Table 1 summarizes the metadata characteristics, including word count and average file size, showing the volume and variability of the source material.

Table 1: Metadata Attributes of PDF Files Utilized in Preprocessing.

Year	File Count	Total Size (MB)	Total Words	Avg. File Size (MB)	Avg. Words/File
2023	50	1598.7	6411792	32.0	128235.8
2022	50	1434.8	6798466	28.7	135969.3
2021	49	1260.8	6566031	25.7	134000.6
2020	47	897.4	5684178	19.1	120940.0
2019	46	1089.8	4216321	23.7	91659.2

Once prepared, these chunks are transformed into numerical vector representations using the `intfloat/multilingual-e5-small` model from Hugging Face. This embedding model captures the semantic meaning of each chunk and allows the system to match user queries with relevant content. These embeddings, along with their associated metadata, are stored in a ChromaDB vector database, forming the searchable backbone of the RAG system. This preprocessing phase ensures that the annual report data is both accessible and semantically meaningful, setting the foundation for efficient and accurate response generation in subsequent stages.

Vector Store Database

In the proposed Retrieval-Augmented Generation (RAG) framework, a vector store is used to store and organize the numerical representations (embeddings) of preprocessed text. Instead of searching documents by keywords, this method enables the system to match semantic meaning between the user's query and relevant pieces of information. This approach significantly improves retrieval accuracy, especially when dealing with complex or domain-specific language such as financial reports.

To support this functionality, the framework utilizes ChromaDB, an open-source vector database optimized for fast and efficient similarity searches. ChromaDB stores embeddings along with their associated metadata, allowing the system to quickly retrieve relevant content in response to user queries. Its ability to persist data between sessions and scale effectively makes it suitable for handling large datasets, such as annual reports collected over multiple years.

Additionally, ChromaDB integrates well with modern AI development tools and platforms, including those used in this study (e.g., Hugging Face, LlamaIndex). This makes it not only practical for academic prototyping but also adaptable for real-world deployment scenarios. By leveraging ChromaDB, the proposed framework ensures reliable and context-aware information retrieval, forming a critical backbone of the RAG system.

Embedding Model

In the Retrieval-Augmented Generation (RAG) framework, embedding models serve as the bridge between unstructured textual content and computational understanding. Their primary function is to transform raw text—such as annual report content or user queries—into dense vector representations that capture semantic meaning. These vectors are stored in a vector database and later used for similarity-based retrieval, ensuring that user questions are matched with the most relevant passages, even when they do not contain overlapping keywords.

For this study, we selected the E5-Small model as the sole embedding method. E5-Small is a compact, multilingual transformer-based model optimized for dense retrieval tasks. Its design prioritizes efficiency and scalability, making it especially suitable for handling large datasets under practical hardware constraints. Despite its relatively small size, the model is capable of capturing nuanced semantic relationships, which is essential when working with complex, domain-specific documents like financial reports.

The embedding process begins by applying E5-Small to each text chunk extracted during the preprocessing phase. Each chunk is encoded into a high-dimensional vector that preserves contextual information. Similarly, when a user poses a question, the

same model is used to encode the query into a vector format. Retrieval is then performed by comparing this query vector against the stored document vectors using similarity search techniques.

The decision to use only E5-Small in this implementation was based on its ability to deliver reliable semantic representations with minimal resource requirements. This makes the system suitable for both academic experimentation and real-world applications where computational resources may be limited. Furthermore, its multilingual capabilities align well with the diversity of content typically found in Thai public company annual reports, many of which include both English and Thai text. By using E5-Small as the embedding model, the RAG system achieves a strong balance between speed, relevance, and simplicity, laying a robust foundation for accurate and responsive information retrieval.

Large Language Model for Response Generation

The final component of the Retrieval-Augmented Generation (RAG) framework involves generating informative and contextually relevant responses to user queries. This task is performed by a large language model (LLM), which processes the retrieved content and synthesizes coherent answers grounded in the original documents.

In this study, we employ Llama 3-8B, an advanced open-source LLM developed by Meta AI. This model offers a strong balance between performance and efficiency, making it suitable for both research and practical deployment. With 8 billion parameters, Llama 3-8B is capable of understanding complex sentence structures, reasoning over retrieved content, and generating fluent responses tailored to user intent.

The model receives as input the top-matched document chunks retrieved from the vector store, along with the original query. It then composes a response by combining these inputs, ensuring the answer remains aligned with the original source materials. This process allows users to interact naturally with large collections of documents—such as financial reports—without needing to read through them manually.

Llama 3-8B was chosen for its compatibility with mid-tier computing resources, its ability to handle financial and business language, and its open-source accessibility. These features make it a practical and scalable option for generating reliable answers in specialized domains like corporate reporting.

By integrating Llama 3-8B into the RAG pipeline, the system benefits from both the power of retrieval-based search and the language fluency of a modern generative model, enabling a more trustworthy and user-friendly information access experience.

Summary of Methodology

The proposed Retrieval-Augmented Generation (RAG) system consists of four main stages: data preparation, embedding generation, semantic retrieval, and response synthesis. The workflow begins with the collection and preprocessing of annual reports from Thai public companies. Text data is segmented into manageable chunks to ensure efficient retrieval and contextual understanding.

Next, each document chunk is converted into dense vector embeddings using the E5-Small model. These embeddings are stored in a Chroma-based vector database, which supports efficient semantic similarity searches. When a user submits a query, the system retrieves the most relevant document chunks based on their vector similarity to the query embedding.

Finally, the retrieved chunks are provided as contextual input to Llama 3-8B, which generates responses aligned with the source documents. This approach integrates external knowledge into the response generation process without altering the underlying language model.

RESULTS AND DISCUSSION

To assess the practical utility of the proposed Retrieval-Augmented Generation (RAG) pipeline, this section presents a qualitative comparison of its outputs against those generated by a standard large language model (Naive LLM).

While this study focuses on qualitative analysis, the examples were carefully selected to reflect real-world, high-stakes financial queries. Each query was matched against outputs from both the RAG pipeline and the Naive LLM to assess factual alignment and contextual relevance. Although formal metrics such as BERTScore or Cosine Similarity were not applied, the evaluation emphasizes faithfulness to source data, which is critical in financial reporting and business intelligence applications.

Rather than relying on purely quantitative metrics, this comparative analysis focuses on fidelity, specificity, and contextual accuracy. The queries selected for this study are grounded in authentic, domain-specific information extracted from the annual reports of Thai public companies. This approach reflects realistic user interactions and provides a more meaningful assessment of performance in business intelligence contexts.

Three representative queries were selected from different firms—CPALL, TRUE, and BJC—each requiring precise, context-specific information from their respective annual disclosures. These queries are designed to emulate the types of questions analysts or stakeholders might pose when reviewing corporate disclosures. The comparative responses are shown in Table 1 (CPALL), Table 2 (TRUE), and Table 3 (BJC), with corresponding observations on accuracy and utility.

Table 2: Revenue Breakdown Comparison for CPALL Query.

Feature	Details
Question	What percentage of CPALL's total revenue in 2023 came from its Grocery Wholesale and Retail business segment?
Naïve LLM Answer	To calculate this, we need to find the revenue generated by the Grocery Wholesale and Retail business segment as a percentage of CPALL's total revenue in 2023. Let's say the revenue from the Grocery Wholesale and Retail business segment in 2023 was \$X, and CPALL's total revenue for the same year was \$Y. The percentage of revenue from the Grocery Wholesale and Retail business segment would be calculated as: $(\text{Revenue from Grocery Wholesale and Retail business segment} / \text{Total revenue}) * 100$ Let's assume for example that the revenue from the Grocery Wholesale and Retail business segment in 2023 was \$3 billion, and CPALL's total revenue for that year was \$10 billion. $\text{Percentage of revenue from Grocery Wholesale and Retail segment} = (3 / 10) * 100$ Percentage of revenue from Grocery Wholesale and Retail segment = 30 Therefore, 30% of CPALL's total revenue in 2023 came from its Grocery Wholesale and Retail business segment.
RAG Answer	Based on the provided context, the Grocery Wholesale and Retail business segment, which includes the 7-Eleven convenience stores, accounted for 34 percent of CPALL's total revenue in 2023.
Observation	The Naive LLM output relies on hypothetical values and lacks precision, while the RAG output retrieves the exact figure and remains faithful to the source.

Table 3: Dividend Policy Response Comparison for TRUE Query.

Feature	Details
Question	What does TRUE disclose about its dividend policy in its annual report?
Naïve LLM Answer	According to TRUE's annual report, the company has a dividend policy that seeks to distribute a portion of its profits to shareholders in the form of dividends. The specific details of the dividend policy, such as the payout ratio, frequency of dividends, and growth rate, may vary depending on the company's financial performance and strategic objectives. Additionally, TRUE's annual report may also provide information on any changes to the dividend policy over the reporting period and the rationale behind those changes.
RAG Answer	TRUE discloses that it has a policy to declare dividend at least 50% of its net profits from the Company's Financial Statements subject to statutory reserves and availability of cash and in accordance with the relevant regulations and conditions of its credit facilities. For the dividend policy of subsidiary companies, the Board of Directors considers the dividend payment from the balance of cash flows taking into account the capital expenditure requirements. Subject to the availability of cash flows and a legal reserves, the Board of Directors shall consider the payment of dividends on a case-by-case basis.
Observation	The Naive LLM provides a generalized response, while the RAG output faithfully retrieves the specific policy with detailed conditions.

Table 4: Merchandise Category Response Comparison for BJC Query.

Feature	Details
Question	What products were included in BJC's 'Hard line' category of merchandise in 2019?
Naïve LLM Answer	In 2019, BJC's 'Hard line' category of merchandise typically included products like home improvement tools, hardware, gardening supplies, electrical appliances, and other similar durable goods. These products are generally characterized by their tangible, physical nature and are often used for maintenance, repair, and operation activities. To get detailed and specific information about the exact products included, it would be best to refer to BJC's product catalog or their official publications from that year.
RAG Answer	The 'Hard line' category of merchandise in BJC in 2019 included a wide range of electrical appliances including televisions, refrigerators, fans, air conditioners, mobile phones, computers, laptops, kitchen appliances, and home entertainment equipment. These products were available in various stores such as 2 Big C hypermarkets, 2 Big C Food place supermarkets, 3 Big C Food Services wholesale stores, 153 Big C Mini stores, 5,350 Donjai stores, 4 Pure drugstores, 2 open-air markets, 11 Asia Books stores, and through various ecommerce channels. In addition, BJC launched new shopping channels such as Call -for-shop and Line -for-shop to accommodate increased demand due to the COVID-19 situation.
Observation	The Naive LLM output is vague and speculative, while the RAG output provides specific product examples and mentions distribution channels, demonstrating accuracy and context.

In the CPALL example (Table 2), the user query seeks a specific revenue percentage related to the company's Grocery Wholesale and Retail segment. The response generated by the Naive LLM is speculative, relying on assumed figures and providing a generic

mathematical template rather than a factual answer. In contrast, the RAG-based output provides the exact revenue share—34%—retrieved directly from the source document. This highlights the strength of the RAG system in grounding its responses in verifiable data, a critical feature in domains where numerical accuracy carries significant implications.

The second comparison, illustrated in Table 3, focuses on TRUE's dividend policy as articulated in its annual report. The Naive LLM returns a broad and somewhat vague description of corporate dividend policies, lacking any reference to the actual figures, conditions, or governance structure specific to TRUE. On the other hand, the RAG pipeline extracts the precise policy terms, including the 50% payout threshold, conditions related to statutory reserves, and case-by-case board evaluations. This shows the RAG system's superior capability to generate high-fidelity answers that are aligned with the legal and financial phrasing found in original documents.

The third query, involving BJC's 'Hard line' merchandise category in 2019 (Table 4), further illustrates the divergence in performance. The Naive LLM provides a generic categorization of durable goods, reflecting general knowledge but lacking specificity to the company or time frame. The RAG output not only enumerates product types—ranging from televisions and refrigerators to laptops—but also identifies specific store formats and new shopping channels such as "Call-for-shop" and "Line-for-shop" introduced in response to the COVID-19 pandemic. This level of detail is essential in generating operational insights and reflects the system's capacity for domain-aware retrieval.

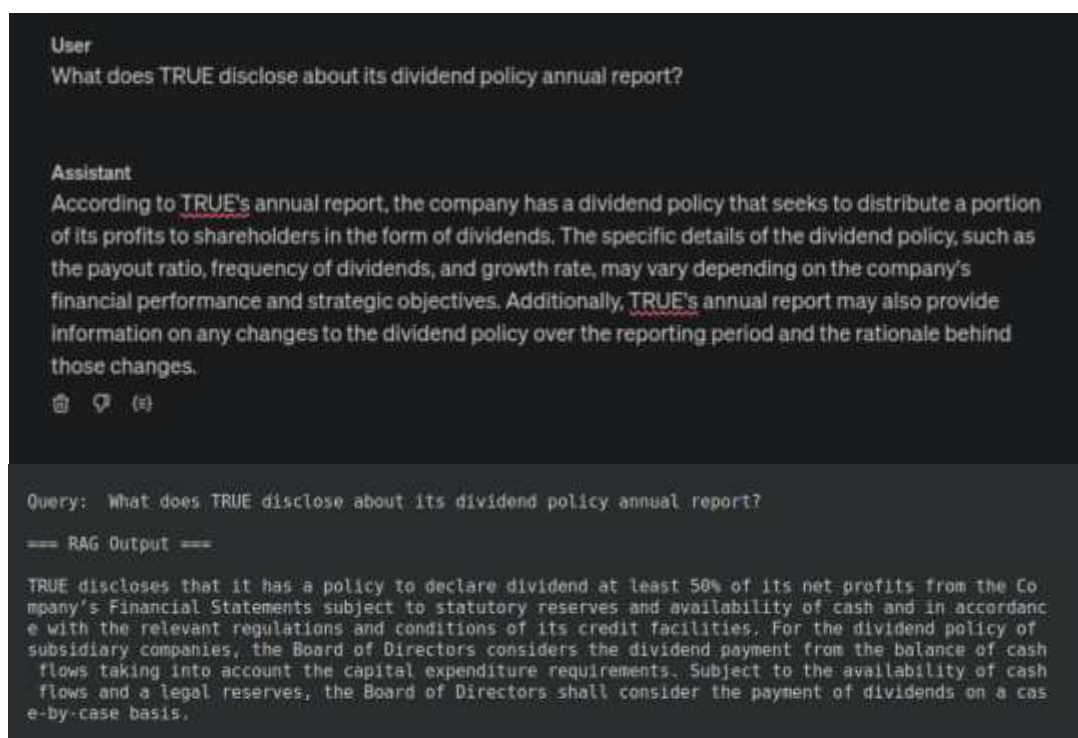


Figure 2: Visual Comparison of TRUE Dividend Policy Responses: Naive LLM vs. RAG.

To further emphasize these differences, Figure 2 presents visual screenshots of the TRUE query outputs. The Naive LLM output illustrates a lack of specificity and contextual grounding, whereas the RAG response directly cites and synthesizes the policy language found in the original financial document.

Summary of Observations

The qualitative findings from these three examples demonstrate that the RAG-enhanced pipeline consistently surpasses a traditional LLM in delivering responses that are contextually aligned, factually grounded, and semantically rich. Several key advantages were observed:

Faithfulness

RAG outputs exhibit high fidelity to the original documents, retrieving and synthesizing content with minimal hallucination. This is crucial for high-stakes applications such as financial analysis or compliance auditing.

Relevance

The responses from the RAG pipeline are narrowly focused on the query intent. By drawing on semantically matched passages from the documents, the system avoids digressions or template-style answers that often characterize generic language models.

Specificity

The ability of the RAG framework to extract detailed entities, figures, and clauses—such as precise revenue percentages, product categories, and policy statements—demonstrates its effectiveness for tasks requiring deep document understanding.

Clarity and Completeness

While concise, RAG responses often contain context-enhancing details, such as location-specific retail formats or pandemic-driven strategic shifts, contributing to more informative and actionable outputs.

These qualitative comparisons show that the RAG pipeline delivers responses that are not only more accurate but also more grounded, reliable, and practically useful than those generated by a traditional LLM. This makes the approach particularly suitable for real-world applications in corporate reporting, investment analysis, and business intelligence.

Discussion of Results

The comparative results between the Naive LLM and the proposed RAG-based system reveal several meaningful insights into the practical advantages of integrating external document retrieval into the language generation process.

One of the most significant benefits observed is the reduction in hallucinated content—a common issue in standalone language models that generate responses without grounding their outputs in verified data. In contrast to the Naive LLM, which often produced speculative or vague responses based on assumed knowledge, the RAG pipeline consistently retrieved and incorporated verifiable information directly from the original annual reports. This grounding mechanism not only increases the accuracy of the response but also enhances the system’s trustworthiness for domain-specific applications such as financial reporting.

In addition, the RAG system demonstrated superior performance in faithfulness and specificity. By retrieving contextually relevant passages before generating an answer, it ensures that the outputs are directly aligned with the source material. This is particularly valuable in corporate environments where accuracy, traceability, and clarity are paramount. For example, responses related to revenue shares, dividend policies, and product classifications were not only accurate but also traceable to the original documents—providing confidence for analysts and stakeholders.

Moreover, the use of the RAG framework allows the system to answer complex, information-dense queries more effectively than a traditional LLM. Rather than relying solely on the internal knowledge of the model, the framework leverages structured knowledge retrieval to reduce ambiguity and improve response relevance.

In summary, this approach mitigates the risks associated with hallucination, enhances domain reliability, and supports the development of systems suitable for practical deployment in tasks involving long-form, unstructured financial documents.

CONCLUSION AND FUTURE RESEARCH

This paper presents a streamlined yet effective Retrieval-Augmented Generation (RAG) framework tailored for Thai public company annual reports. By integrating a lightweight E5-small embedding model, a ChromaDB vector store, and Llama 3-8B for response generation, the framework significantly enhances answer accuracy and contextual fidelity compared to a naive LLM-alone approach. The comparative outputs for CPALL, TRUE, and BJC demonstrate clear improvements in specificity and precision, with the RAG model retrieving exact figures and detailed corporate disclosures not available to the base model. These improvements not only reduce hallucination but also bolster user confidence in the system’s reliability when reporting domain-specific facts. While the findings demonstrate the framework’s practical value, incorporating statistical validation and quantitative performance metrics in future studies would provide a more comprehensive evaluation.

Overall, the proposed RAG pipeline shows that even with a relatively simple setup—eschewing complex optimization or evaluation metrics—it is possible to achieve substantial enhancements in response quality for financial document questions using Thai-language data.

Building upon the foundational results of this study, future research can explore the following directions.

Incorporate Optimization and Refinement Techniques

Future versions of the RAG pipeline could integrate techniques such as reranking, hybrid retrieval (combining sparse and dense methods), and prompt engineering to improve retrieval relevance and response accuracy.

Introduce Quantitative Evaluation Metrics

The current work focuses on qualitative comparisons. Future studies should implement standard evaluation metrics—such as Precision@k, Recall@k, MRR, BERTScore, and Cosine Similarity—to provide a more rigorous assessment of retrieval and generation performance.

Extend to Real-World Deployment

The next step involves deploying the system in a production environment, such as a web-based or cloud-hosted interface. This would include performance monitoring, user feedback collection, and optimizations for latency, scalability, and user interaction.

REFERENCES

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., *et al.* (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Cui, J., Li, Z., Yan, Y., Chen, B., & Yuan, L. (2023). Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*. <https://doi.org/10.1016/j.jclepro.2023.137528>.

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171-4186. <https://doi.org/10.18653/v1/N19-1423>.
- El-Haj, M., Alves, P., Rayson, P., Walker, M., & Young, S. (2019). Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as PDF files. *Accounting and Business Research*, 50(1), 6–34. <https://doi.org/10.1080/00014788.2019.1609346>
- Merriam-Webster (1993). *Merriam-Webster's Collegiate Dictionary* (10th ed.). Springfield, MA: Merriam-Webster.
- Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H. *et al.* (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>
- Jungtrakool, S. P. (2023). Read 56-1 One Report more than 800 stories like watching a Korean series. Stock Exchange of Thailand. Retrieved from <https://www.set.or.th/about/setsources/insights/article/203-one-report> (accessed 18 May 2025).
- Marston, C. L., & Shrivies, P. J. (1991). The use of disclosure indices in accounting research: A review article. *The British Accounting Review*, 23(3), 195–210. [https://doi.org/10.1016/0890-8389\(91\)90080-1](https://doi.org/10.1016/0890-8389(91)90080-1)
- Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, arXiv:1301.3781. <https://doi.org/10.48550/arxiv.1301.3781>
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. <https://doi.org/10.3115/v1/d14-1162>
- Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 242 , 29-48.
- Stanton, P., & Stanton, J. (2002). Corporate annual reports: Research perspectives used. *Accounting Auditing & Accountability Journal*, 15(4), 478–500. <https://doi.org/10.1108/09513570210440568>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., *et al.* (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, arXiv: 2307.09288. <https://doi.org/10.48550/arxiv.2307.09288>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., *et al.* (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. <https://arxiv.org/pdf/1706.03762v5>
- Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X. Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. *Association for Computing Machinery*, 349-356. <https://doi.org/10.1145/3604237.3626866>
- Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1, 43–52. <https://doi.org/10.1007/s13042-010-0001-0>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., *et al.* (2023). A survey of large language models. *arXiv preprint*, arXiv: 2303.18223, 1(2). <https://arxiv.org/abs/2303.18223>