The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 4, No. 1. Online ISSN: 2831-4794

EconVNNewsBot-2: Enhancing Economic News Retrieval with Retrieval-Augmented Thoughts for Improved Question-Answering in Vietnam

Dang-Khoa Nguyen-Le[a], Tuyet Hue Tran[a,b], Minh-Chau Pham[*,c], Hanh Thi Hong Tu[d], Tan-Cong Nguyen[d]

[a]Faculty of Information Technology, University of Science - VNUHCM, Ho Chi Minh City, Vietnam; [b]School of Business Information Technology, University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam; [c]AI Lab - AI VIETNAM, Ninh Thuan, Vietnam, [d]University of Social Sciences and Humanities, VNU-HCM, Ho Chi Minh City, Vietnam

*chaupm.bn@gmail.com; phone 84 903292973

ABSTRACT

The rapid expansion of Vietnamese economic news presents challenges for investors and entrepreneurs seeking accurate, domain-specific information. Traditional search engines struggle with complex queries, revealing a gap in effective information retrieval. In our previous work (Paper ID: 22 in ATAC 2024), we introduced EconVNNewsBot, a specialized Question-Answering (QA) system for Vietnamese economic news, where we used Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) reasoning separately. Although effective, these methods treated retrieval and reasoning as distinct steps, limiting system efficiency and coherence. This paper extends our prior work by employing a more advanced approach: Retrieval-Augmented Thoughts (RAT). RAT unifies retrieval and reasoning into a single integrated framework, providing more precise and context-sensitive responses with improved efficiency. Our new experiments show that the RAT-based approach enhances user engagement, reduces retrieval time, and simplifies the information-gathering process. These findings suggest that RAT can transform how economic news is accessed, benefiting both information seekers and news agencies. The research is supported by the unique EconVNNews dataset, compiled using the EconVNNewsCrawl tool developed by our team, ensuring the system's robust performance in the economic domain.

Keywords: Vietnamese economic news, Question-Answering (QA) system, Retrieval-Augmented Thoughts (RAT), Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT) reasoning, Semantic chunking.

1. INTRODUCTION

In modern journalism, the goal is not only to deliver information but also to ensure that readers can easily comprehend and engage with the content [1], [2]. Accessing accurate, timely, and domain-specific information is critical, especially in economics, where the volume of daily articles can overwhelm readers. Investors, entrepreneurs, and decisionmakers often struggle to quickly locate precise, relevant information using traditional search engines like Google or Bing. Despite improvements, these engines still struggle to deliver fast, accurate responses to complex, domain-specific queries.

A promising solution to this challenge lies in Question-

Answering (QA) systems, which can process natural language questions and automatically retrieve precise answers from both structured and unstructured documents [3]. QA systems are particularly beneficial in the economic domain, where accurate and timely information is vital for decision-making.

This paper presents EconVNNewsBot-2, an upgraded version of our previous QA framework for Vietnamese economic news. Unlike its predecessor, which combined Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) reasoning separately, EconVNNewsBot-2 integrates both retrieval and reasoning into a unified framework using the novel Retrieval-Augmented Thoughts (RAT) approach. By combining these components, RAT enables more coherent, context-aware answers, improving both accuracy and efficiency. EconVNNewsBot-2 reduces search time, enhances response accuracy, and improves user interaction with economic news content, offering an efficient tool for readers and newsrooms alike.

The main contributions of this study include:

- Applying the **Retrieval-Augmented Thoughts (RAT)** approach to economic domain, which unifies retrieval and reasoning for more effective, context-sensitive question answering.
- Developing **EconVNNewsBot-2**[1] – an upgraded automated chatbot that answers readers' questions related to economic news, leveraging the RAT framework for enhanced performance.

The remainder of this paper is organized as follows: Section II covers related work, Section III explains the article collection process and the architecture of EconVNNewsBot-2, Section IV presents the dataset, Section V discusses and analyzes the results, and Section VI summarizes the conclusions.

2. LITERATURE REVIEW

The integration of advanced natural language processing (NLP) techniques in question-answering (QA) systems has revolutionized how information is accessed and interpreted across various domains. This section examines the evolution of methodologies such as Retrieval-Augmented Generation (RAG) [4], Chain-of-Thought (CoT) reasoning [5], and their recent amalgamation in Retrieval Augmented Thoughts (RAT). We explore their implications for enhancing QA systems, particularly in low-resource language contexts such as Vietnamese economic news [6].

Retrieval-Augmented Generation (RAG) represents a significant breakthrough in QA systems by bridging the gap between retrieval and generation processes. By leveraging both parametric memory (e.g., a pre-trained transformer model) and non-parametric memory (e.g., vectorized external documents), RAG improves the factual grounding of generated responses [4], [7]. Its dual memory setup ensures that QA systems can retrieve relevant documents and generate accurate, contextually relevant answers. This approach has proven particularly effective in domains requiring precision, such as finance, healthcare, and law [8], [9]. Despite its successes, RAG faces limitations when applied to complex multi-step reasoning tasks. Traditional retrieval techniques often fail to predict the relevance of information for subsequent reasoning steps, leading to errors in generation. This shortcoming is especially pronounced in knowledge-intensive applications, where precise retrieval of intermediate information is critical.

Chain-of-Thought (CoT) Reasoning enhances the logical coherence of QA systems by decomposing complex queries into intermediate reasoning steps [5]. This approach mimics human cognitive processes, allowing models to solve multistep problems by iteratively building upon previous conclusions. CoT has demonstrated significant improvements in tasks like mathematical reasoning and multi-hop question answering. However, CoT's reliance on the initial query and its intermediate steps often introduces hallucinations—factually incorrect or logically flawed reasoning—due to the model's inability to access external knowledge dynamically. This limitation restricts its applicability in domains requiring rigorous factual accuracy.

Retrieval-Augmented Thoughts (RAT) synergizes the strengths of RAG and CoT, addressing their respective limitations by introducing an iterative process for refining reasoning steps with external knowledge. Unlike traditional RAG, which retrieves information in a single pass, RAT iteratively revises each intermediate step by querying a knowledge base with context derived from both the task and prior reasoning [10]. This progressive refinement significantly mitigates hallucination, ensuring that each reasoning step is contextually grounded and factually accurate. RAT's causal reasoning approach—where intermediate steps influence the retrieval and generation of subsequent steps—aligns closely with human problem-solving methods. Studies have shown that RAT outperforms standalone RAG and CoT techniques in benchmarks for code generation, mathematical reasoning, and creative writing, achieving notable improvements in accuracy and coherence [10].

Semantic Chunking is another critical technique in improving QA performance, especially in long or multi-topic texts. It involves breaking down input text into semantically meaningful units, or "chunks," that allow the model to focus on relevant segments, improving the retrieval and generation processes [11]–[13]. This method reduces computational complexity and enhances accuracy, particularly in domains like law, medicine, and finance. It has proven effective in improving document retrieval and answer generation in complex, multi-topic queries [14].

Embedding Techniques convert text into dense vectors that capture the semantic meaning of the content, allowing for effective similarity measurements between queries and answers. Popular models like BERT and XLM-R have advanced text embedding for English and other high-resource languages [15], [16]. For Vietnamese, PhoBERT, a pre-trained model,

[1] https://github.com/KhoaLearn/EconVNNewsBot-2

facilitates effective Vietnamese text embedding [17], while newer multilingual models, such as Jina-Embeddings-v3, offer further advancements in text embedding for low-resource languages like Vietnamese [18].

Developing QA systems for low-resource languages, such as Vietnamese, presents unique challenges, including limited annotated datasets and linguistic diversity [19]. By iteratively refining intermediate reasoning steps, RAT has the potential to address the gaps in low-resource language QA systems. Its ability to dynamically retrieve and contextualize information makes it particularly suited for domains such as Vietnamese economic news, where precision and contextual understanding are paramount. EconVNNewsBot-2, the subject of this study, incorporates RAT to address these specific challenges in retrieving and processing Vietnamese economic news, providing more accurate, context-aware answers while reducing inefficiencies present in previous systems.

3. PROPOSED METHOD

3.1 Semantic Chunking

The EconVNNewsBot-2 uses the o200k_base tokenizer [20] to tokenize the content of articles and the jina-embeddingsv3 model [18] via JinaAI's Segmenter-API to divide the text into chunks based on semantic relationships [21]. This method begins with representing the input text T as a sequence of tokens: $T = \{t_1, t_2, \ldots, t_n\}$.

The goal is to identify chunks: $C = \{c_1, c_2, \ldots, c_m\}$ where each chunk c_i encapsulates a distinct semantic meaning. Tokens are converted into a vector embedding $v_t = E(t_i)$ using pre-trained models which measure semantic similarity between tokens using metrics like cosine similarity. Chunk boundaries are determined by the jina-embeddings-v3 model [18] to predict chunk labels based on context $P(c_i \mid context)$ [21].

3.2 Embedding Model Jina-Embedding-V3

The EconVNNewsBot-2 uses the jina-embeddings-v3 model, designed for multilingual data and long-context tasks, with 570 million parameters and support for up to 8192 tokens [18]. It incorporates LoRA (Low-Rank Adaptation) adapters [22] for tasks like document retrieval and ranking. Built on the Transformer architecture and the Jina-XLMRoBERTa framework, it replaces absolute with relative positional encoding to capture word dependencies [16]. The multi-head attention mechanism allows the model to grasp complex semantic relationships in long texts. The embedding vector of a text sequence X is computed as follows: $fv_{\text{jina}} = f\text{jina}(x1, x2, \ldots, xn)$.

Where: f_{jina} is the embedding function of the *jinaai/jinaembeddings-v3* model. $x_1, x_2, \ldots, x_n$ are the words in the text sequence. $\mathbf{v}_{\text{jina}}$ is the resulting embedding vector that represents the entire text sequence.

Each embedding vector $\mathbf{v}_{\text{jina}}$ captures the semantic meaning of the entire text sequence, making it efficient for querying and retrieval in the *EconVNNewsBot-2* system.

The jina-embeddings-v3 model ranks among the top multilingual models in the MTEB (Massive Text Embedding Benchmark), performing well for languages like Vietnamese [23]. It was chosen for its lightweight power and ranks in the top 30, as shown in Figure 1.

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
19	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.4
20	text-embedding-004	1200	4.47	768	2048	66.31
21	TDTE					65.96
22	GritLM-8x7B	46703	173.98	4096	32768	65.66
23	jina-embeddings-v3	572	2.13	1024	8194	65.51
24	gte-large-en-v1.5	434	1.62	1024	8192	65.39
25	LLM2Vec-Meta-Llama-3-supervis	7505	27.96	4096	8192	65.01
26	cde-small-v1	143	0.53	768	512	65
27	LLM2Vec-Mistral-supervised	7111	26.49	4096	32768	64.8
28	echo-mistral-7b-instruct-last	7111	26.49	4096	32768	64.68

Figure 1. MTEB - Massive Text Embedding Benchmark [23], [24]

3.3 Vector Database

The system uses Pinecone [25], a vector database to store the embedding vectors. When a user submits a query q, it is embedded into a vector $e_q = f_{\text{model}}(q)$. The system then searches for relevant articles by comparing e_q with the stored article vectors e_i. The query process minimizes the cosine distance between vectors:

$$\min(\cos(e_q, e_i)), \quad \forall i.$$

The output returns the most semantically similar articles, ensuring the most relevant news is retrieved and presented to the user.

3.4 Re-Ranking Model

Table I compares model performance across benchmarks, showing that the jina-reranker-v2-base-multilingual model used in EconVNNewsBot-2 excels across tasks, making it well-suited for this application.

Table 1. Performance comparison of different models across various benchmarks [26].

Benchmark	JinaV2	BgeV2	mMiniLMv2	JinaV1
Model Size	278M	568M	118M	137M
MKQA (26 langs)	54.83	54.17	53.37	-
BEIR (17 datasets)	53.17	53.65	45.40	52.45
MLDR (13 langs)	68.95	59.73	28.91	-
CodeSearchNet (3 tasks)	71.36	62.86	51.78	-
AirBench (zh/en)	61.33	61.28	56.46	-
ToolBench (3 tasks)	77.75	78.46	58.39	74.13
TableSearch	93.31	74.86	53.60	72.89

After retrieving N articles with the highest similarity from the Pinecone, the *jina-reranker-v2-base-multilingual* [26], [27] model re-ranks them to select the top K most relevant articles, enhancing accuracy. This model processes sentence pairs (query and article title/content), embedding and evaluating them for relevance using a semantic classification mechanism. The relevance score for each article A_i is based on the similarity between the query q and the article's content. This process can be expressed as: score(q, A_i) = reranker(q, A_i). Where q is the user's query. A_i represents the i-th article, combining its title and content into a single text. *reranker* is the re-ranking function of the *jina-reranker-v2-base-multilingual* model.

Before re-ranking, the system retrieves N articles from Pinecone based on embedding vector matching. The re-ranking model then refines these by reassessing their relevance to the query, selecting the top K (where $K \leq N$) based on the highest scores. The final K articles are determined using the following formula:

$$K = \underset{A_i \in N}{\text{argmax}}\ \text{score}(q, A_i), \quad \text{for} \quad i = 1,2,...,N. \qquad (1)$$

Where N is the set of articles retrieved from Pinecone. K represents the final set of selected articles based on the highest re-ranking scores. score(q, A_i) is the relevance score of article A_i with respect to the query q.

The re-ranking process filters out irrelevant results and prioritizes articles most semantically relevant to the query, significantly improving accuracy and ensuring the *EconVNNewsBot-2* provides the most valuable information.

3.5 Retrieval-Augmented Thoughts (RAT)

The Retrieval Augmented Thoughts (RAT) framework enhances question-answering (QA) systems by iteratively combining retrieval and reasoning processes. Unlike traditional methods that handle retrieval and reasoning in isolation or a linear sequence, RAT incorporates dynamic refinement at each step, ensuring reasoning is supported by relevant and accurate external knowledge. The following steps outline the RAT framework:

1) Initial Chain-of-Thought Drafting:
   - Input: A user-provided task prompt I.
   - Output: An initial zero-shot reasoning chain $T_1, T_2, ..., T_n$ generated using a large language model (LLM).
   - This reasoning chain is a step-by-step outline of the solution but may contain flaws or incomplete logic.
2) Initial Retrieval:
   - Input: The task prompt I and the initial reasoning chain $T_1, T_2, ..., T_n$.
   - Process: Relevant information is retrieved from a knowledge base or document repository for each step T_i, denoted as R_i.
   - Output: A set of retrieved context documents $R_1, R_2, ..., R_n$ supporting the reasoning steps.
3) Iterative Revision of Chain-of-Thought:
   - Input: Current reasoning step T_i, its retrieved context R_i, and all previous refined steps $T_1^*, T_2^*, \ldots, T_{i-1}^*$.
   - Process: Each reasoning step T_i is revised using its corresponding context R_i to produce an updated reasoning step T_i^*.
   - Output: A refined reasoning chain $T_1^*, T_2^*, \ldots, T_n^*$ that incorporates retrieved knowledge iteratively.
4) Final Answer Synthesis:
   - Input: The complete refined reasoning chain $T_{1:n}^*$.
   - Process: Insights from the refined reasoning chain are synthesized into a comprehensive final answer A.
   - Output: A validated final answer A, cross-checked with retrieved data to ensure accuracy and coherence.
5) Result Presentation:
   - Input: The final answer A and sources $S_1, S_2, ..., S_m$. • Process: Present the final answer to the user with a transparent explanation of each reasoning step and references to the sources.
   - Output: A complete, traceable response that includes reasoning, retrieved evidence, and source attributions.

This structured, iterative process ensures that RAT achieves higher accuracy and contextual relevance compared to traditional methods by continuously refining both reasoning and retrieval processes.

4. DATASET AND SYSTEM OVERVIEW

4.1 EconVNNews Dataset

1) EconVNNews: A Vietnamese Economic News Dataset: The EconVNNews dataset is a comprehensive collection of over 268,000 articles covering economic events, market analyses, and expert opinions on macroeconomic trends in Vietnam from September 2020 to September 2024. This dataset is invaluable for economic research and NLP applications, including question-answering, trend analysis, and market forecasting.

To create EconVNNews, we developed the EconVNNewsCrawl tool, designed to collect data from a variety of reputable Vietnamese news sources, ensuring the dataset is both diverse and reliable. The tool supports the following key features:

- Multiple reputable news sources: EconVNNewsCrawl collects articles from prominent news outlets, such as CafeBiz, CafeF, VietNamNet, and VnExpress.
- Flexible data collection period: Users can specify custom timeframes for data collection, making the system adaptable to time-specific research needs.

- JSON data structure: The collected articles are stored in a structured JSON format, with fields such as "title", "URL", "content", "published date", "author", "category", and "source", enabling easy integration with various analysis tools and platforms.

Together, EconVNNews and EconVNNewsCrawl provide a robust foundation for developing and testing economic NLP models, including those used in question-answering systems like EconVNNewsBot-2.

Table 2 illustrates a steady increase in article numbers, with notable growth between 2021 and 2023, reflecting Vietnam's post-pandemic economic recovery. The sharp rise in 20222023 corresponds to the volatility in financial markets, while the slight drop in 2024 can probably be attributed to the dataset's coverage up to September and potential stabilization in the market.

Table 2. Number and proportion of articles by year and source.

Source	2020	2021	2022	2023	2024	Total
Number of Articles						
cafebiz	572	2190	2798	4524	4247	**14331**
cafef	13192	43882	47996	52598	40590	**198258**
vietnamnet	2610	7933	7060	6351	3990	**27944**
vnexpress	2567	8258	6368	6199	4800	**28192**
Total	**18941**	**62263**	**64222**	**69672**	**53627**	**268725**
Proportion (%)						
cafebiz	0.21	0.81	1.04	1.68	1.58	**5.33**
cafef	4.91	16.33	17.86	19.57	15.10	**73.78**
vietnamnet	0.97	2.95	2.63	2.36	1.48	**10.40**
vnexpress	0.96	3.08	2.37	2.31	1.79	**10.49**
Total	**7.05**	**23.17**	**23.90**	**25.93**	**19.96**	**100.00**

In conclusion, the EconVNNews dataset effectively captures fluctuations in the economic market and media trends during Vietnam's post-pandemic recovery, making it a valuable resource for analyzing the country's economy.

2) Reliability of Information: To ensure the reliability of the data, the information collection process follows these criteria:

- Verified sources: Each article is accompanied by author details, publication date, and a link to a reputable source. • Accuracy: Data is drawn from official financial reports and recognized economic institutions.
- Expert analysis: Many articles feature insights from economic experts, which further enhance the data's quality and utility.

By adhering to these standards, the EconVNNews dataset delivers dependable and precise information, supporting research and data-driven decision making.

4.2 System Overview

EconVNNewsBot-2 is designed to automatically answer user questions about Vietnamese economic news by combining data preprocessing, retrieval-augmented reasoning, and iterative chain of thought (CoT) refinement. As illustrated in Figure 3, the system consists of three main phases:

1) Data Preprocessing: The EconVNNewsCrawl tool collects reputable Vietnamese economic news articles, storing them in a structured JSON dataset. Each article is broken into manageable segments via semantic chunking, and the system extracts relevant metadata (e.g., titles, publication dates, URLs). Next, each chunk is converted into a vector representation (using a language-model embedding) a stored in a vector database. This preparation ensures that the system can later retrieve content based on semantic similarity.

2) Retrieval-Augmented Thoughts (RAT): When a user asks a question, EconVNNewsBot-2 uses a retrieval module to retrieve the most relevant chunks from the vector database. These chunks are then passed to an LLM Agent that iteratively refines its reasoning using chain-of-thought prompts. In each iteration (*"Step 1 ... Step n"* in the figure), the agent augments the retrieved segments with intermediate thoughts, re-queries the vector database if needed, and updates its reasoning. This iterative cycle of retrieval and refinement, which we term *Augmented Revision*, helps the agent to handle complex questions with more accuracy.

3) Answer Generation: After reaching a stable and well supported chain of thought, the LLM agent composes a final response. The system appends relevant references, such as article URLs or excerpted passages, making the response both traceable and transparent. This answer is then returned to the user, concluding the QA workflow.

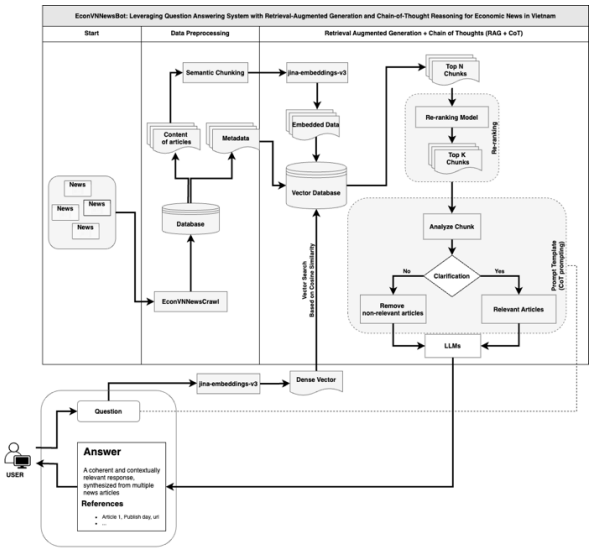


Figure 2. System Architecture of EconVNNewsBot (RAG+CoT)

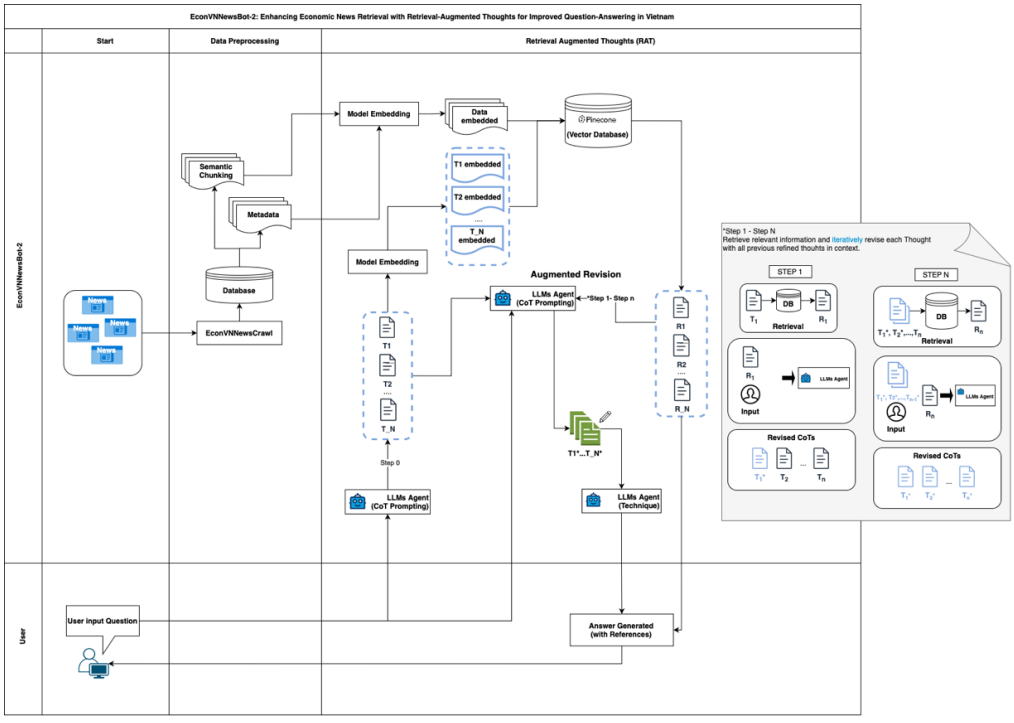


Figure 3. System Architecture of EconVNNewsBot-2 (RAT)

5. EMPIRICAL ANALYSIS

This section presents a detailed empirical analysis of the Retrieval-Augmented Thoughts (RAT) framework (Figure 3) compared to the Retrieval-Augmented Generation (RAG) with Chain-of-Thought (CoT) reasoning (Figure 2). The analysis evaluates the models across three dimensions: (1) retrieval and evidence usage, (2) reasoning transparency and controllability, and (3) empirical performance of response time.

5.1 Retrieval and Evidence Usage

The RAT framework demonstrates an improved ability to systematically retrieve and organize evidence. Specifically, RAT integrates a structured Chain-of-Thought (CoT) reasoning mechanism, allowing for hierarchical problem decomposition. For instance, when addressing the query, "Which bank performed best in 2022?", the RAT system systematically identified relevant criteria such as profitability, customer satisfaction, and operational efficiency. It then utilized reputable sources, including The Asian Banker and Global Banking and Finance, to justify its recommendations.

In contrast, RAG+CoT aggregated information from diverse sources but lacked a structured reasoning mechanism. The final conclusion, e.g., "TPBank was the best-performing bank in 2022," was not directly supported by a detailed evaluation of specific performance metrics, reducing the interpretability of its evidence chain.

Example Output (RAT):

- Techcombank's accolades (e.g., "Best Digital Bank for Customers" by Global Finance).
- ACB's recognition for risk management by The Asian

  Banker.
- Supporting links directly to evaluated sources.

Example Output (RAG+CoT):

- Identified TPBank as the best-performing bank based on aggregated rankings, but the evidence lacked detailed metric-based comparisons.

5.2 Reasoning Transparency and Controllability

RAT emphasizes reasoning transparency by explicitly linking retrieved documents to each step in its Chain-of-Thought reasoning. The framework ensures controllability by enabling researchers or users to track how specific evidence contributes to the final answer.

For example, RAT's output for the banking query clearly articulated intermediate steps, such as defining evaluation criteria (e.g., profitability, customer preference), contextualizing geographical scope, and referencing relevant awards. This structured reasoning enhances user trust and facilitates error analysis.

RAG+CoT, while effective at summarizing multiple sources, did not articulate intermediate reasoning steps, making it challenging to evaluate the reliability of its conclusions or intervene in its reasoning process.

5.3 Empirical Performance of Response Time.

Table 3 provides a summary of the empirical performance of response time of the evaluated frameworks. The Retrieval-Augmented Thoughts (RAT) framework demonstrates substantial improvements in response times compared to RAG+CoT while maintaining competitive quality of the answer, outperforming other model in handling complex queries efficiently.

Table 3. Models Performance of Response Time (seconds)[2].

Category	EconVNNewsBot	EconVNNewsBot-2
Fact-Based Question	79.8	13.1
Analytical/Inference-Based Question	111.4	29.7

[2] https://github.com/KhoaLearn/EconVNNewsBot-2

Time-Specific Question	75.91	17.06
Comparative Question	103.69	19.03

EconVNNewsBot-2 demonstrates substantial improvements in response times across all categories, showcasing its optimization over the original EconVNNewsBot. These results highlight the system's capability to address a wide spectrum of queries efficiently, supporting its role in advanced economic news retrieval and analysis.

5.4 Implications for Real-World Applications

The findings suggest that RAT is better suited for tasks requiring high reasoning transparency and multi-step problem-solving, such as financial analysis or academic research. In contrast, RAG+CoT is more appropriate for applications where summarization is prioritized over in-depth reasoning.

6. DISCUSSION

This study presents a paradigm shift in the development of QA systems tailored to domain-specific challenges, focusing on Vietnamese economic news. The introduction of the Retrieval-Augmented Thoughts (RAT) framework marks a significant innovation by seamlessly integrating retrieval and reasoning processes into a unified and iterative model. This approach addresses the inherent limitations of earlier methodologies, such as Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) reasoning, which often treated retrieval and reasoning as discrete steps, leading to inefficiencies and gaps in contextual alignment.

Through a preliminary empirical analysis, the superiority of RAT is clearly demonstrated. Unlike traditional models, RAT dynamically refines its reasoning process at each step by incorporating relevant external evidence, ensuring that responses are precise, contextually accurate, and logically coherent. This iterative refinement mitigates common pitfalls like hallucination and incomplete logic, especially in scenarios requiring complex, multi-step reasoning. The framework's dynamic retrieval capability ensures that every sub-task is well supported by pertinent information, making it exceptionally suited for the fast-paced and intricate domain of economic news.

The EconVNNews dataset, a foundational component of this research, further enhances the capabilities of the system. This meticulously curated dataset encompasses over 268,000 articles from reputable Vietnamese news outlets, offering a rich repository for the model to operate effectively. Advanced techniques such as semantic chunking and embedding models, specifically jina-embeddings-v3, have been employed to structure and process this data, enabling the RAT framework to extract and utilize context with unparalleled efficiency.

The practical implications of this research are profound. EconVNNewsBot-2, powered by RAT, sets a new approach for QA systems in low-resource languages. It transforms how stakeholders - such as investors, entrepreneurs, and policymakers - access and interpret economic information. By providing accurate, transparent, and context-aware responses, the system not only enhances decision-making but also fosters trust and engagement among users.

This research also highlights the broader significance of unifying retrieval and reasoning as a principle for future QA systems. It underscores how advancements in natural language processing can directly address real-world challenges, offering a replicable and scalable model for other domains and languages.

7. CONCLUSION

The Retrieval-Augmented Thoughts (RAT) framework represents a groundbreaking contribution to the evolution of QA systems, offering a dynamic and iterative approach that unifies retrieval and reasoning processes. The application of RAT in EconVNNewsBot-2 demonstrates its potential to address the intricate and fast-changing nature of Vietnamese economic news, supported by the rich and well-structured EconVNNews dataset. This system provides a critical tool for improving access to reliable, actionable information in Vietnam's rapidly developing market.

Vietnam's economic landscape, as an emerging market, is characterized by its rapid growth, increasing global integration, and the unique challenges of navigating a dynamic information environment. Tools like EconVNNewsBot-2 play a vital role in supporting informed decision-making in such contexts. By providing accurate and context-sensitive insights, these systems empower stakeholders to better understand market trends, mitigate risks, and seize opportunities, thereby contributing to the broader development of Vietnam's economy.

The insights gained from this research highlight the potential for similar systems to drive innovation in other emerging markets, where access to precise and timely information is critical. Future work could expand the RAT framework to multilingual or cross-domain applications, ensuring adaptability to diverse market needs. Additionally, incorporating user feedback and improving scalability could enhance its relevance and impact, further solidifying its role as a foundational tool for specialized information retrieval in fast-evolving economies like Vietnam's.

REFERENCES

[1] M. Mencher and W. P. Shilton, "News reporting and writing: Brown & benchmark publishers", *United Kingdom*, 1997.

[2] G. Starkey, *Balance and bias in journalism: Representation, regulation and democracy*. Bloomsbury Publishing, 2017.

[3] M. A. Calijorne Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems", *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 6, pp. 635–646, 2020.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktaschel¨ *et al.*, "Retrieval augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[6] P. Tran, D. Nguyen, H.-A. Tran, T. Nguyen, and T. Tran, "Building a closed-domain question answering system for a low-resource language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 3, pp. 1–14, 2023.

[7] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.

[8] S. Setty, K. Jijo, E. Chung, and N. Vidra, "Improving retrieval for rag based question answering models on financial documents," *arXiv preprint arXiv:2404.07221*, 2024.

[9] R. Kalra, Z. Wu, A. Gulley, A. Hilliard, X. Guan, A. Koshiyama, and P. Treleaven, "Hypa-rag: A hybrid parameter adaptive retrieval augmented generation system for ai legal and policy applications," *arXiv preprint arXiv:2409.09046*, 2024.

[10] Z. Wang, A. Liu, H. Lin, J. Li, X. Ma, and Y. Liang, "Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation," *arXiv preprint arXiv:2403.05313*, 2024.

[11] FullStackRetrieval-com, "5 levels of text splitting," https://github.com /FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/Leve lsOfTextSplitting/5_Levels_Of_Text_Splitting.ipynb, 2024, [Accessed 19-10-2024].

[12] LlamaIndex, "Semantic chunker - llamaindex — docs.llamaindex.ai," https://docs.llamaindex.ai/en/stable/examples/node_parsers/semantic_chunking/, 2024, [Accessed 19-10-2024].

[13] LangChain, "How to split text based on semantic similarity | langchain — python.langchain.com," https://python.langchain.com/v0.2/docs/how_to/semantic-chunker/, 2024, [Accessed 19-10-2024].

[14] "Semantic Chunking for RAG: Better Context, Better Results — multimodal.dev," https://www.multimodal.dev/post/semantic-chunking-for-rag, [Accessed 19-10-2024].

[15] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] A. Conneau, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[17] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," *arXiv preprint arXiv:2003.00744*, 2020.

[18] S. Sturua, I. Mohr, M. K. Akram, M. Gunther, B. Wang, M. Krimmel, F. Wang, G. Mastrapas, A. Koukounas, N. Wang *et al.*, "jinaembeddings-v3: Multilingual embeddings with task lora," *arXiv preprint arXiv:2409.10173*, 2024.

[19] K. Nguyen, V. Nguyen, A. Nguyen, and N. Nguyen, "A Vietnamese dataset for evaluating machine reading comprehension," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2595–2605. [Online]. Available: https://aclanthology.org/2020.coling-main.233

[20] OpenAI, "GitHub - openai/tiktoken: tiktoken is a fast BPE tokeniser for use with OpenAI's models. — github.com," https://github.com/openai/tiktoken, 2024, [Accessed 19-10-2024].

[21] JinaAI, "Segmenter API — jina.ai," https://jina.ai/segmenter, 2024, [Accessed 19-10-2024].

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[23] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022.

[24] N. Muennighoff, N. Tazi, and . Magne, "MTEB Leaderboard - a Hugging Face Space by mteb — huggingface.co," https://huggingface.co/spaces/mteb/leaderboard, 2022, [Accessed 20-10-2024].

[25] E. Liberty, "The vector database to build knowledgeable ai | Pinecone — pinecone.io," https://www.pinecone.io/, 2019, [Accessed 11-10-2024].

[26] J. AI, "jina-reranker-v2-base-multilingual," 2023, accessed: 20-102024, available at Jina Reranker Model. [Online]. Available: https: //huggingface.co/jinaai/jina-reranker-v2-base-multilingual

[27] S. Ognawala, J. Fu, Y. Zhang, and S. Martens, "Jina Reranker v2 for Agentic RAG: Ultra-Fast, Multilingual, Function-Calling & Code Search — jina.ai," https://jina.ai/news/jina-reranker-v2-for-agentic-rag-ultra-fast-multilingual-function-calling-and-code-search/, 2024, [Accessed 20-10-2024].