Thesis for Bachelor's Degree

A Hybrid Search RAG Chatbot Service for Accounting and Tax Domains

Doyoon Song

Department of Electrical Engineering and Computer Science

Gwangju Institute of Science and Technology

2025

학사학위논문

회계, 세무 업계를 위한 하이브리드 검색 기반 RAG 챗봇

송도윤

전기전자컴퓨터공학부

광주과학기술원

2025

A Hybrid Search RAG Chatbot Service for Accounting and Tax Domains

Advisor: Sundong Kim

by

Doyoon Song

Department of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology

A thesis submitted to the faculty of the Gwangju Institute of Science and Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in the Electrical Engineering and Computer Science Concentration

Gwangju, Republic of Korea
May 27, 2025
Approved by

Professor Sundong Kim
Committee Chair

A Hybrid Search RAG Chatbot Service for Accounting and Tax Domains

Doyoon Song

Accepted in partial fulfillment of the requirements for
the degree of Bachelor of Science

May 27, 2025

Committee Chair _____
Prof. Sundong Kim

Committee Member _____
Prof. Mansu Kim

Dedicated to my family.

BS/EC 20185091 Doyoon Song. A Hybrid Search RAG Chatbot Service for Accounting and Tax Domains. Department of Electrical Engineering and Computer Science. 2025. 15p. Advisor: Prof. Sundong Kim.

Abstract

In this paper, I propose a hybrid Retrieval-Augmented Generation (RAG) chatbot system specifically designed for the accounting and tax domains. In the domain of accounting and tax, accessing accurate and reliable legal information such as laws, internal regulations, and organizational bylaws requires both semantic understanding and lexical relevance. However, traditional retrieval-based methods often fail to capture the nuanced meaning of these documents, leading to ambiguous or low-quality responses, especially in cases involving vague or incomplete queries. The system combines lexical and semantic search via a hybrid Vespa-based search engine and leverages a multi-agent structure to categorize and refine user intents before retrieval. Additionally, pre-processing strategies are implemented to build vertical hierarchies across unstructured data and are deployed with prompt engineering techniques to handle inappropriate or overly ambiguous queries. The service has shown robust performance even in edge cases with over 5,000 unique users and 12,000+ message interactions. The service demonstrates high usability and reliability, providing accurate, hierarchical, and reference-backed answers in a sensitive and legally complex domain. The chat-bot

service is available at Accounting and Tax Chat-bot.

BS/EC 20185091 송도윤. 회계, 세무 업계를 위한 하이브리드 검색 기반 RAG 챗봇. 전기전자컴퓨터공학부. 2025. 15p. 지도교수: 김선동.

국 문 요 약

본 논문에서는 회계 및 세무 분야에서 좋은 성능을 내도록 설계된 하이브리드 검색 기반 생성(Retrieval-Augmented Generation, RAG) 챗봇 시스템을 제안한다. 회계 및 세무 분야에서는 자연어 기반 검색 질의를 처리하는 과정에서 법령, 내규, 조직 규정 등과 같은 법적 정보를 높은 정확도로 검색하기 위해 의미적론적 해석과 문자 기반 유사도 검색이 모두 요구된다. 그러나 기존의 의미 기반 검색은 이러한 문서들의 미묘한 의미를 충분히 포착하지 못해, 모호한 사용자 질의의 경우 품질이 낮은 응답을 초래하는 문제가 있다. 이 시스템은 Vespa 기반의 하이브리드 검색 엔진을 활용하여 의미 기반 검색과 문자 기반 유사도 검색을 결합하여 사용하고, 검색 전에 사용자의 질의를 분류하고 정제하기 위해 다중 에이전트 구조를 도입한다. 더불어, 비정형 데이터 간 위계 구조를 사전에 구축하고, 모호하거나 부적절한 질의를 처리하기 위해 프롬프트 엔지니어링 기법을 적용하였다. 이 시스템은 다양한 엣지 케이스에서도 견고한 성능을 보였으며, 누적 5,000명 이상의 누적 사용자와 12,000건 이상의 메시지 상호작용을 기록하였다. 이를 통해 민감하고 법적 복잡성이 높은 도메인에서도 정확하고 계층적인 구조를 갖춘 레퍼런스 기반 답변을 안정적으로 제공할 수 있음을 입증하였다. 해당 챗봇 서비스는 Accounting and Tax Chat-bot 에서 확인할 수 있다.

Contents

List of Figures

Chapter 1

Introduction

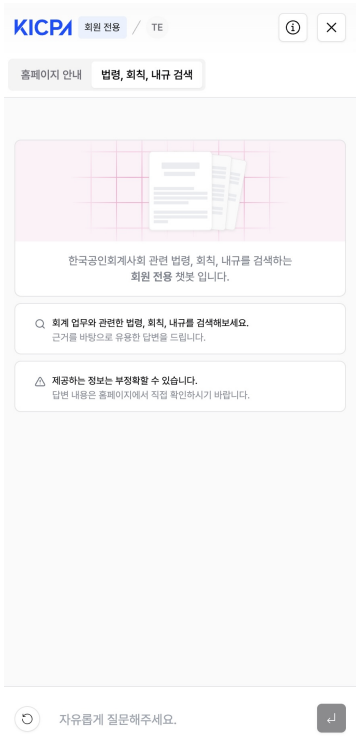

Figure 1.1: The Accounting and Tax Chat-bot that are deployed currently at Accounting and Tax Chat-bot for CPAs. It shows robust performance on user queries in the domain of accounting and tax.

Difficulty in Semantic Inference for Legal and Regulatory Documents Legal texts such as laws, internal regulations, and organizational bylaws often rely heavily on referencing other documents to establish their meaning and authority. In the accounting and tax domains, these documents frequently cite specific articles, clauses, or external legal bases to define their scope. However, naive semantic search methods

typically fail to capture this inter-document linkage, resulting in poor retrieval performance. Without incorporating the relational structure between documents, semantic inference remains shallow and often misleading. Therefore, retrieval systems in this domain must account not only for semantic similarity but also for the underlying citation and dependency structure, which often is best captured through lexical search in order to achieve accurate and contextually relevant results.

Ambiguity in Semantic Categorization of Accounting and Tax Domain Data Data in the accounting and tax domain often lack clear semantic categorization, making it difficult to distinguish the intended context of a given document. For example, a document labeled as "tax agency form materials" could refer either to procedural information or to downloadable form templates, depending on its content and usage. This semantic ambiguity causes retrieval systems to surface documents that may be topically related but fail to match the user's actual intent producing vague or "gray" responses. To address this challenge, we implement a multi-agent architecture that first classifies the user query into a predefined semantic category. This category then guides the retrieval process, enabling the system to filter and prioritize content based on both the alignment of meaning and intent.

Lack of Hierarchical Relationships Among Data In many accounting and tax information systems, documents are often related through implicit hierarchies for example, a guide document may refer to an entire bulletin board of certain website, while individual posts within that board may provide more specific guidance. When a user's

query does not directly match any individual post, it would be natural for the system to fall back to a broader level document or relevant category. However, the absence of explicitly modeled hierarchical relationships between these entities prevents such behavior. To address this, we introduce a pre-processing pipeline that establishes hierarchical links between documents and higher level structures, enabling the system to support fallback behaviors and navigate content across multiple levels of granularity.

Performance Degradation Due to Abusive or Incomplete Queries Typical retrieval systems are also highly vulnerable to vague, incomplete, or abusive user queries. In domains dealing with sensitive or regulated topics, such failures are not just inconvenient but potentially harmful. Systems must not only attempt to refine or clarify such queries but also have the ability to tactfully refuse responses to inappropriate or high-risk questions when necessary.

To address these challenges, we propose a hybrid Retrieval-Augmented Generation (RAG) chat-bot system specifically designed for the accounting and tax domains. Our system combines lexical and semantic search strategies, uses a multi-agent structure to categorize and refine user queries, establishes vertical hierarchies among data, and implements prompt engineering to handle inappropriate or vague queries. This paper details the design, implementation, and impact of the system in a sensitive and legally complex environment.

Chapter 2
Functionalites

2.1 Category Selection

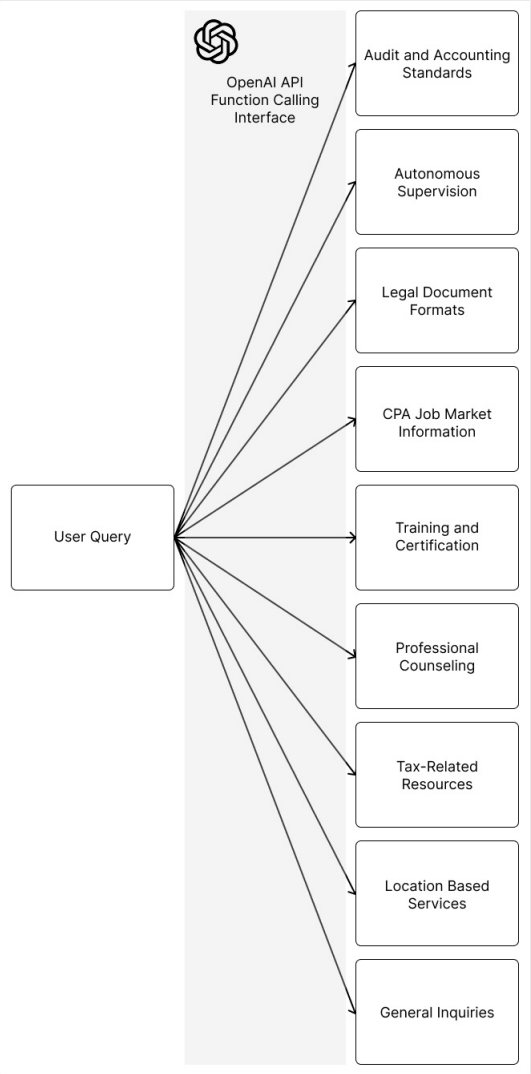

Figure 2.1: Multi-Agent Architecture to filter adequate category for given user query with OpenAI API function calling feature.

To further enhance the alignment between user intent and retrieved content, the system performs semantic categorization of user queries prior to retrieval. This is achieved

using OpenAI's GPT-4o mini model with the function calling feature. By invoking a structured function call interface, the model infers the most appropriate category for a given query based on its semantics rather than relying solely on keyword matching. This category prediction is then used to guide the retrieval process, allowing the system to narrow the document search space and improve the contextual relevance of the returned results.

The predicted categories correspond to core functional areas of the Certified Public Accountants, such as audit and accounting standards, autonomous supervision, legal document formats, CPA job market information, training and certification, professional counseling, tax-related resources, location-based services, and general inquiries. Each category is linked to a specific parameter, enabling the model to semantically route user queries to the most appropriate content domain. This structured classification ensures that even highly contextual or specialized queries are directed to the right source of information.

2.2 Hybrid Semantic-Lexical Search

To address the limitations of naive semantic search in interpreting accounting and tax domain data, our system implements a hybrid retrieval architecture that combines both semantic and lexical strategies. Specifically, it leverages a multi-stage query rewriting and retrieval pipeline to improve performance in handling inter-document dependencies. The core retrieval functionality is exposed via the `POST /search` endpoint, which processes queries from users alongside optional metadata such as category or directory scope.

Document Search The request includes the raw query string, a category indicator (optional), and the directory target.

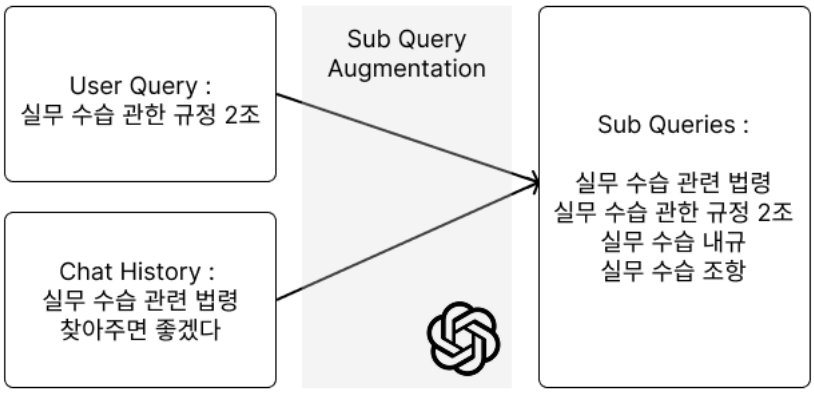

Figure 2.2: Sub-Query Augmentation Process, the specified category is omitted for simplicity.

Sub-query Augmentation Complex or multi-part queries are decomposed into sub-queries targeting specific statutes or regulations using OpenAI's structured output capabilities. In particular, certain documents rely on the citation of other documents to achieve contextual completeness, and sub-queries are augmented to retrieve these cited or referencing materials accordingly.

Context-Aware Query Restructuring : Past chat history is used to clarify or expand the current query.

Vector Store Retrieval The rewritten query is embedded and used to retrieve semantically similar documents within the specified category.

Hierarchical Retrieval Fallback In cases where the primary retrieval step fails to surface a document with sufficient content to fully address the user's query intent, our system employs a hierarchical fallback mechanism. This mechanism escalates the search to a broader scope by retrieving high-level bulletin or board-style documents that comprehensively cover related topics. These documents serve as thematic hubs and often contain aggregated explanations, procedural guides, and links to more specific resources. By surfacing these general but topically relevant documents, the system ensures that users are still provided with meaningful guidance even when an exact

match is unavailable. This approach improves recall without compromising the contextual relevance of responses, and is particularly effective in domains where semantic gaps and sparse labeling are common.

Pruning and Fetching Candidate documents are filtered using OpenAI's schema-based structured outputs to eliminate irrelevant content. Document IDs are used to fetch actual content from an S3-based storage layer.

Deduplication Retrieved documents with identical titles are merged to reduce redundancy.

2.3 Data Pre-processing in Retrieval through Document Linkage and Structured Fallback

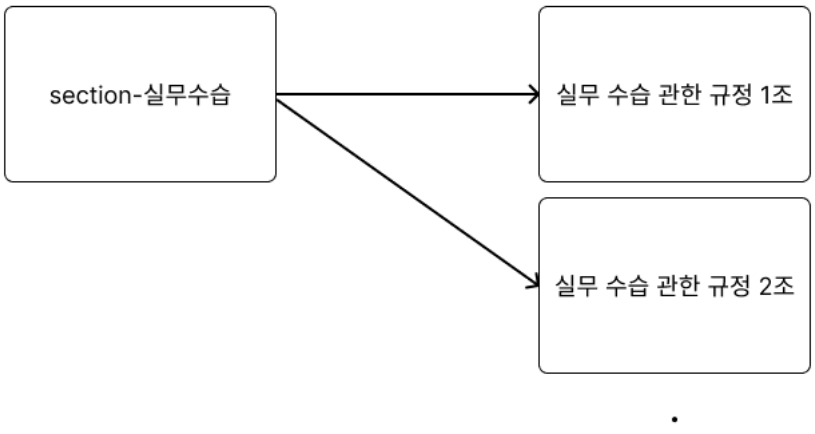

Figure 2.3: Fallback Retrieval Data Structure, high-level data entries are prefixed with "section," enabling the system to supply broader contextual information when specific documents do not adequately address the user's query.

To address the limitations posed by the lack of explicit hierarchical relationships in accounting and tax-related document collections, we introduce a data pre-processing pipeline that enriches retrieval performance through structural awareness. Specifically, we isolate and manage bulletin board-level data separately by assigning a distinct "sec-

tion" prefix to their titles, enabling clear distinction from individual post-level documents. This structuring allows the system to recognize when a user's query is better matched at a broader level of granularity. In cases where no direct document match is found, the system performs a structured fallback by retrieving only the most semantically relevant sections (i.e., bulletin boards) rather than irrelevant or noisy individual results. This approach not only improves retrieval accuracy but also enhances user experience by providing higher-level context and guiding users toward more focused exploration.

2.4 Mitigating Retrieval Risks from Abusive or Incomplete Queries

To enhance robustness in handling incomplete, sensitive, or otherwise problematic user queries, we introduce a series of prompt engineering and query rewriting strategies. First, to prevent the system from engaging with potentially controversial topics—such as the overlapping scope between CPAs and tax agents—we explicitly restrict responses to such queries through instruction-level prompt control. Additionally, we mitigate performance degradation caused by unnatural spacing patterns by normalizing whitespace during pre-processing. Domain-specific jargon frequently used in the accounting field is also addressed through synonym rewriting, allowing the system to better align queries with formal terminology in the indexed corpus. Lastly, previous user interactions are incorporated into the current query context to preserve conversational continuity and reduce semantic drift across turns, especially in multi-step information-seeking dialogues.

Query Rewriting *Whitespace Rewriter* Removes unnecessary spaces based on a predefined threshold. *Synonym Rewriter* Substitutes synonyms using a domain-specific dictionary to improve semantic alignment with lexical parsing strategy.

Chapter 3
System Design

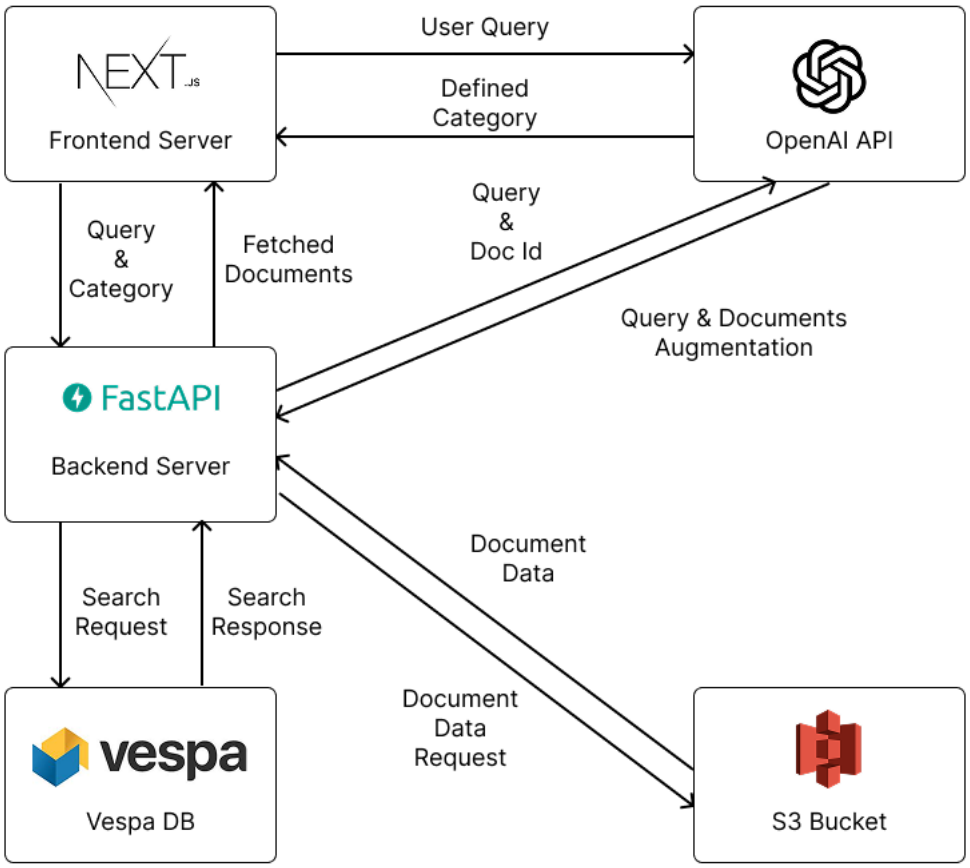

Figure 3.1: System Architecture of chat-bot system.

3.1 System Architecture

Figure 3.1 illustrates the end-to-end architecture of the hybrid retrieval system. When a user submits a query through the Next.js frontend, the system first invokes the OpenAI API to infer the most appropriate category using a function-calling interface. This defined category is then returned to the frontend and passed, along with the original query, to the FastAPI-based backend server. The backend issues a search

request to the Vespa DB using both the query and the category information to retrieve semantically and contextually relevant document IDs. Additionally, when sub-query augmentation is required, the OpenAI API is invoked again with both the original query and retrieved document data to improve query specificity. Based on the final set of document IDs, the backend retrieves detailed document data from an S3 bucket and returns the fully resolved result set to the frontend for presentation. This modular, category-guided architecture allows for context-aware retrieval, improved query specificity, and fallback handling through integration across multiple services.

3.2 Category Selection

When a user submits a query, the system forwards it to the OpenAI GPT-4o mini model, which uses function calling to predict the most appropriate category. If the model successfully infers a specific category, the query is returned accordingly, enabling targeted retrieval from the relevant content domain. In cases where no category is confidently predicted, the system defaults to a broader or fallback search strategy. This conditional branching ensures that user intent is semantically interpreted and mapped to domain-specific knowledge boundaries, thereby improving both retrieval precision and system robustness.

3.3 Hybrid Semantic-Lexical Search

Figure 3.1 corresponds to the hybrid retrieval architecture implemented to address the limitations of naive semantic search in the accounting and tax domain. The process begins with a user submitting a query via the Next.js frontend, which is then sent to the OpenAI API to infer the appropriate category using function calling. The frontend passes the query along with the inferred category to the FastAPI backend server, which coordinates the multi-stage retrieval pipeline. At this point, complex or multi-part queries are decomposed into sub-queries via OpenAI's structured output capabilities to account for inter-document dependencies, such as citation-based contextual relationships.

These sub-queries are embedded and issued to the Vespa vector store for semantic retrieval. Retrieved candidates are pruned through schema-based classification to en-

sure contextual and legal relevance. In parallel, a hierarchical retrieval logic is applied to resolve section-level fallbacks when no matching document is found. The resulting document IDs are then used to fetch full document content from an S3-based storage layer. To reduce redundancy, documents with identical titles are deduplicated before being returned to the frontend. This architecture enables precise, legally grounded search across both public and member-restricted content by integrating semantic inference, lexical fallback, and structured post-processing in a cohesive pipeline.

3.4 Mitigating Retrieval Risks from Abusive or Incomplete Queries

To ensure safe and reliable retrieval in a sensitive professional domain, the system implements a series of query rewriting and prompt engineering techniques designed to mitigate the risks posed by incomplete, ambiguous, or inappropriate queries. At the pre-processing stage, a Whitespace Rewriter normalizes unnatural spacing patterns that can hinder tokenization and matching. A Synonym Rewriter replaces informal or domain-specific jargon with standardized terms based on a curated accounting dictionary, improving alignment with both semantic and lexical retrieval strategies. For controversial queries such as those involving disputes over the legal scope between CPAs and tax agents the system leverages instruction-level prompt control to tactfully avoid generating potentially inappropriate responses. Additionally, the system maintains context continuity by incorporating past chat history into current queries, allowing for more coherent multi-turn interactions. These combined strategies enable the system to not only handle edge cases gracefully but also deliver responses that are contextually sound, regulation-aware, and aligned with user intent.

Chapter 4

Impact

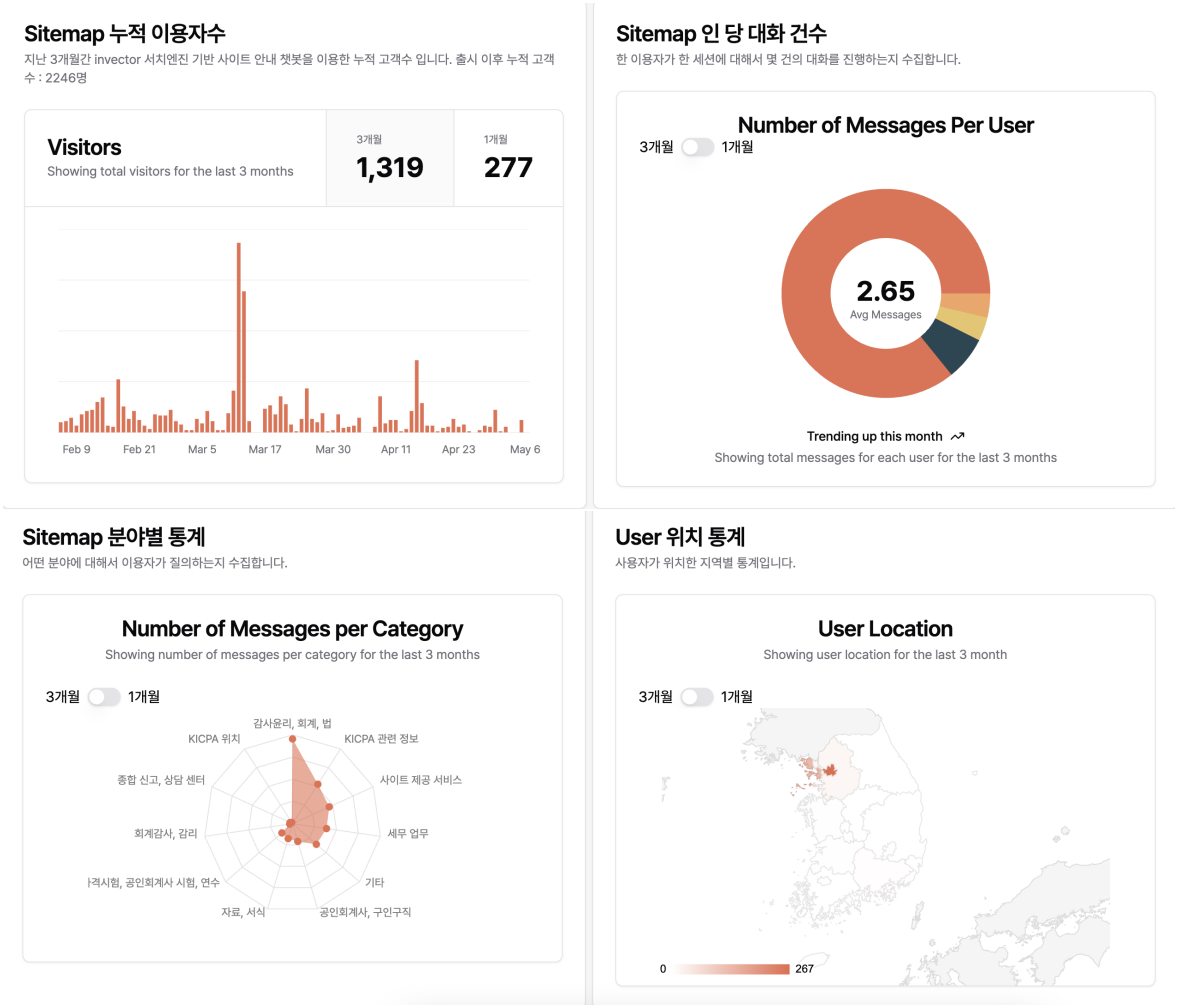

Figure 4.1: Impact and Analytics

4.1 Impact and User Interactions

Since its official launch in late December 2024, the chat-bot service has been actively operating and serving both certified accountants and general users. To date, the system has supported over unique 5,000 users and accumulated more than 12,000 messages across all sessions. In the past three months alone, more than 2,500 unique users have

interacted with the service, indicating sustained and growing engagement. Although continuous session retention was not explicitly measured, daily usage patterns suggest that over 200 accountants actively utilize the chat-bot on a regular basis. On average, each session comprises approximately 2.65 messages per user, reflecting meaningful engagement in information-seeking interactions.

User conversations primarily center around professional and domain-specific topics, reflecting the practical value of the chatbot service. The most frequently discussed subjects include audit ethics, accounting, and law—highlighting users' demand for accurate and reliable guidance in complex regulatory areas. This is followed by inquiries related to the Korean Institute of Certified Public Accountants (KICPA) and its official website, as well as tax-related operations. Other common themes include the CPA examination, job opportunities, document templates and forms, qualification tests, audit reviews, integrated tax filing, and support center services. This distribution of topics suggests that the chat-bot serves not only as a domain-specific knowledge assistant but also as a practical gateway to essential administrative and procedural information.

User location data indicates that the majority of interactions originated from major metropolitan and regional areas in South Korea. The highest concentration of users came from Seoul, followed by Incheon, Gyeonggi-do, Gyeongsangnam-do, Chungcheongnam-do, and Jeju Island. This geographic distribution suggests strong adoption in urban centers, while also demonstrating meaningful engagement from users in southern and coastal regions, reflecting the nationwide relevance and accessibility of the service.

Also, as a result of deploying the chatbot, KICPA achieved a 30% reduction in incoming phone inquiries and successfully maintained over average of 50 daily active users. This indicates not only a measurable decrease in manual support demands but also consistent user engagement with the chatbot service, demonstrating its effectiveness in addressing user needs and improving operational efficiency.

Chapter 5
Conclusion

This study presents the design, deployment, and successful operation of a hybrid Retrieval-Augmented Generation (RAG) chatbot tailored for the accounting and tax domains. By combining semantic categorization, structured fallback mechanisms, and query rewriting strategies, the system demonstrates robust performance in handling domain-specific queries while securing steady user engagement from thousands of professionals and general users nationwide. The chatbot's architecture and retrieval pipeline have proven effective in delivering accurate, context-aware responses in a high-stakes, regulation-heavy environment.

Nevertheless, there remain areas for improvement. When legal and general-purpose documents are mixed without clear boundaries, retrieval performance tends to degrade. Additionally, the current monorepo structure used to deploy separate chat-bot versions has revealed limitations in maintainability, as the divergence in business logic complicates updates and testing. Furthermore, the existing category schema was partially shaped by developer assumptions, which introduces subjectivity. Future iterations could benefit from a more data-driven, inductive approach to category definition in order to enhance classification accuracy and retrieval precision.

References

1. OpenAI, "Openai api reference." https://platform.openai.com/docs/api-reference/introduction, 2024. Accessed: 2025-05-06.

2. V. Team, "Vespa documentation." https://docs.vespa.ai/, 2024. Accessed: 2025-05-06.

Acknowledgements

I am sincerely grateful to Professor Sundong Kim for his insightful guidance and generous support throughout this study, as well as for his encouragement and advice as I explored various career paths. I would also like to thank my peers and friends for their constant encouragement, thoughtful feedback, and companionship throughout this journey.