VIETTEL DIGITAL SERVICES

AI Agent for Q&A, Customer Care, and Proactive Sales on ViettelPay Pro app

Viettel Digital Talent 2025

Mentee:
Minh An Nguyen

Mentor:
Do Khac Phong
Lam Xuan Thu

June 2025

Contents

Abstract

I developed an AI agent system for automated customer support on ViettelPay Pro, a Vietnamese fintech application, addressing the critical need for simultaneous, accurate responses to user inquiries in the Vietnamese language. The system implements Anthropic's contextual retrieval methodology combined with LangGraph-based agent orchestration to handle complex multi-turn conversations, error resolution, and procedure guidance.

My approach centers on a contextual knowledge base built from business documentation, where I enhanced document chunks with LLM-generated context before embedding. The system uses an ensemble retrieval strategy combining BM25 with Vietnamese tokenization and ChromaDB with specialized Vietnamese embeddings (`dangvantuan/vietnamese-document-embedding`), followed by Cohere reranking. A LangGraph workflow manages conversation state and routes queries through intent classification, query enhancement, and response generation nodes.

I conducted comprehensive evaluation using synthetic datasets generated with GPT-4.1 while using Gemini 2.0 Flash for system responses to prevent self-enhancement bias. The results demonstrate strong performance: single-turn retrieval achieved 98.0% Hit@10 and 93.2% Hit@5 with 80.1% MRR, while multi-turn retrieval maintained 89.8% Hit@10 and 86.3% Hit@5 with 68.5% MRR. Intent classification reached 80.4% overall accuracy with 87.8% recall for critical error-help intents and 96.9% recall for human escalation requests.

The system provides a scalable foundation for Vietnamese customer support automation, demonstrating that contextual retrieval can significantly improve response quality in domain-specific applications while maintaining conversation coherence across multiple turns.

1 Introduction

ViettelPay Pro, a specialized application for Viettel agents and retail points, requires an AI solution capable of handling diverse support scenarios from basic FAQ responses to complex error troubleshooting. The business requirements for this system extend beyond simple question-answering. Users need guidance through multi-step financial procedures, consistent error resolution scripts, and the ability to escalate to human agents when necessary. The challenge is compounded by the Vietnamese language's unique characteristics, including compound words, context-dependent meanings, and domain-specific financial terminology that standard NLP models may not handle effectively.

I designed and implemented an AI agent that addresses these challenges through three key innovations. First, I applied Anthropic's contextual retrieval methodology to enhance knowledge base construction, where each document chunk receives LLM-generated context before embedding with vietnamese-document-embedding of dangvantuan (2019), significantly improving retrieval accuracy (Anthropic, 2024). Second, I developed a LangGraph-based workflow that manages conversation state across multiple turns while routing queries through specialized processing nodes based on intent classification. Third, I created a

comprehensive evaluation framework using cross-LLM validation to ensure unbiased performance assessment.

The system architecture integrates several components: ChromaDB for semantic search with Vietnamese embeddings, BM25 for keyword-based retrieval with Vietnamese tokenization, Cohere for result reranking, and Gemini 2.0 Flash for natural language generation. The LangGraph orchestration manages conversation memory, intent classification, query enhancement, and response generation through a state machine that adapts to different interaction patterns.

My contribution extends beyond implementation to methodological rigor in evaluation. I generated synthetic datasets using GPT-4.1 while evaluating system performance with Gemini 2.0 Flash, preventing the self-enhancement bias common in LLM-based systems. The evaluation covers single-turn retrieval, multi-turn conversation handling, and intent classification across seven categories relevant to customer support scenarios.

The results validate the effectiveness of contextual retrieval for Vietnamese customer support applications. The system achieved near-perfect information retrieval (98.0% Hit@10) while maintaining strong performance in complex multi-turn scenarios (89.8% Hit@10). Intent classification demonstrates reliable categorization of user requests with particularly strong performance on business-critical intents like error resolution and human escalation.

This work demonstrates that careful adaptation of recent advances in retrieval-augmented generation can create production-ready customer support systems for specialized domains and languages, providing a template for similar applications in Vietnamese fintech and beyond.

A live demo of my system is available at: https://huggingface.co/spaces/minhan6559/viettelpay-chatbot

2 Business Requirements and Data Analysis

2.0.1 Domain Requirements

ViettelPay Pro requires AI support across three critical areas. Error resolution handles over 30 distinct error codes from system timeouts (W02, W04) to payment failures (606, 974), each requiring specific troubleshooting steps and customer communication protocols. Procedure guidance covers complex multi-step processes for mobile top-up, bill payment, and transaction cancellations that vary by network provider, payment method, and transaction context. Policy information encompasses fee structures, transaction limits, and compliance procedures that change based on user classification and transaction type.

Vietnamese language complexity adds challenges through mixed technical-vernacular terminology, agent abbreviations ("gd" for giao dịch, "tk" for tài khoản), and context-dependent meanings where terms shift based on conversation flow.

2.0.2 Data Sources

I built the knowledge base from three sources. The primary document is a Word file contains 14,847 characters of structured procedures with 33 error code mapping tables, step-by-step guides with conditional branches, and policy definitions with numeric thresholds. This serves as a main document to ingest to the knowledge base of the agent. There is also an Excel file provides 9 question-answer pairs demonstrating expected conversational tone and response format. Another Excel file contains 8 predefined interaction scenarios for greetings, escalations, and conversation management.

2.0.3 Content Complexity

The documentation challenges traditional retrieval through several characteristics. Tables dominate error handling with complex code-to-solution mappings. Procedures follow conditional logic where next steps depend on previous outcomes - transaction cancellations vary by timing, type, and verification success. Policy information spans multiple dimensions with commission structures varying by transaction volume and regulatory constraints differing by customer status.

Cross-references create additional complexity where error procedures reference fee policies and transaction guides link to cancellation regulations. I identified 180+ distinct information units requiring individual accessibility while maintaining contextual relationships for complex multi-step guidance scenarios.

3 System Architecture

3.1 Overall Architecture Design

I designed the system around a LangGraph-based agent orchestration that manages conversation state while coordinating specialized processing nodes. The architecture separates concerns into four primary layers: contextual knowledge management implementing Anthropic's retrieval methodology, conversation state management with multi-turn memory, intent-driven workflow routing, and Vietnamese-optimized natural language processing.

The knowledge layer implements Anthropic's contextual retrieval where I enhance document chunks with LLM-generated context before embedding. An ensemble retrieval strategy combines BM25 with Vietnamese tokenization and ChromaDB with specialized embeddings, followed by Cohere reranking for result optimization. The conversation layer maintains state through `ViettelPayState` objects that track message history, processing context, and workflow metadata across multiple turns using InMemorySaver with thread-based management.

The workflow layer uses LangGraph's state machine to route queries through intent classification, query enhancement, knowledge retrieval, and response generation nodes based on conversation context and user intent. The language processing layer handles Vietnamese-specific challenges through custom tokenization, domain-term preservation, and culturally appropriate response generation.

3.2 Contextual Knowledge Base Implementation

Following Anthropic's contextual retrieval methodology (Anthropic, 2024), I implemented a sophisticated document processing pipeline that transforms traditional RAG into contextual RAG. The core insight from Anthropic's research is that individual chunks often lack sufficient context for accurate retrieval, particularly when queries use different terminology than the chunk content but relate to the same concepts when viewed within the broader document context.

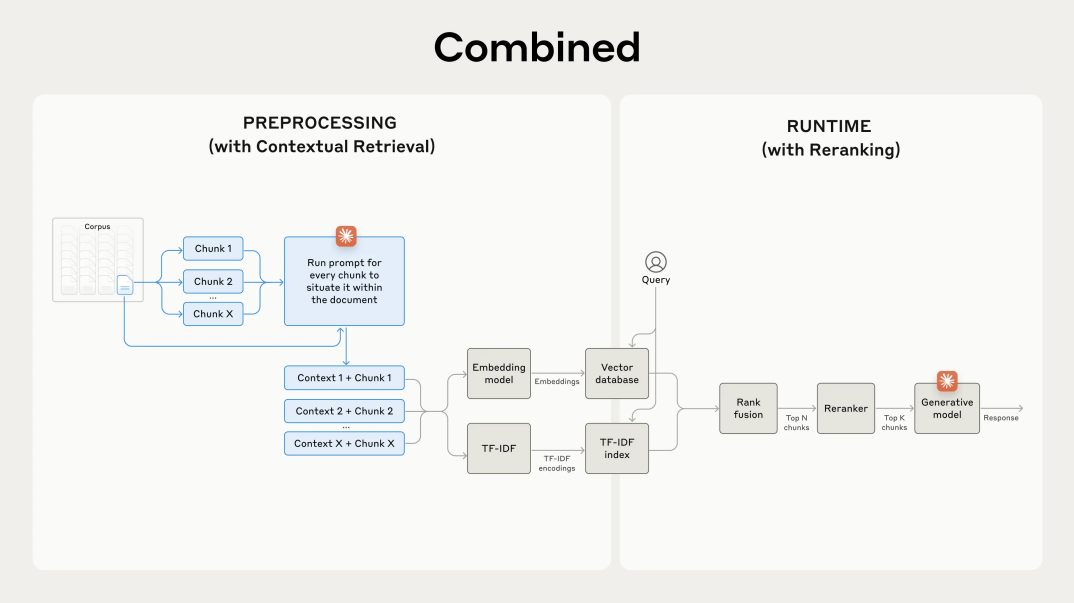

Figure 1: Contextual Retrieval Pipeline

3.2.1 Automated Word Document Processing

I developed the `ContextualWordProcessor` to eliminate manual document preparation. The automated system uses `python-docx` for precise document structure extraction and `markitdown` to capture complete document content for context generation.

The processor automatically detects document hierarchy through heading analysis, supporting both styled headings (`Heading 1`, `Heading 2`) and manual patterns (`#`, `##`, numbered sections). I maintain section hierarchy throughout processing, ensuring each chunk understands its position within the document structure. Tables receive special handling where each row becomes an individual document with structured metadata, preserving column headers and row relationships.

The extraction process creates comprehensive metadata for each chunk including section titles, hierarchy paths, content types, and original content preservation. This metadata enables sophisticated retrieval strategies where users can find information through section-level queries or specific detail searches.

3.2.2 Contextual Enhancement Process

I implemented Anthropic's contextual retrieval technique with Vietnamese-specific adaptations. For each extracted chunk, I generate contextual information using OpenAI's GPT-4o-mini with a carefully crafted Vietnamese prompt template:

```
<tài_liệu>{WHOLE_DOCUMENT}</tài_liệu>
<đoạn_văn>{CHUNK_CONTENT}</đoạn_văn>
Hãy cung cấp ngữ cảnh và tóm tắt ngắn gọn để giúp định vị đoạn văn này trong toàn bộ tài liệu ViettelPay Pro, nhằm cải thiện khả năng tìm kiếm thông tin. Chỉ trả lời bằng ngữ cảnh ngắn gọn, không cần giải thích thêm.
```

I leverage prompt caching by placing the entire document in the system prompt, significantly reducing costs when processing multiple chunks from the same document. Each chunk only requires sending its specific content, avoiding repetition of the full document context. The generated context provides semantic bridging between the chunk content and related concepts throughout the document.

For example, a search query "lỗi 606" can now match chunks with enhanced context like: "Đoạn văn này nằm trong phần 'Hướng dẫn xử lý một số lỗi thường gặp' của tài liệu ViettelPay Pro. Cụ thể, nó thuộc danh mục 'Danh mục bảng mã lỗi' và cung cấp thông tin về mã lỗi 606, liên quan đến giao dịch thất bại do hệ thống nâng cấp."

I combine the generated context with original content before embedding, creating documents that contain both specific procedural details and broader semantic understanding. This addresses the fundamental limitation where embedding models may miss important relationships between chunks that are clear when viewing the complete document.

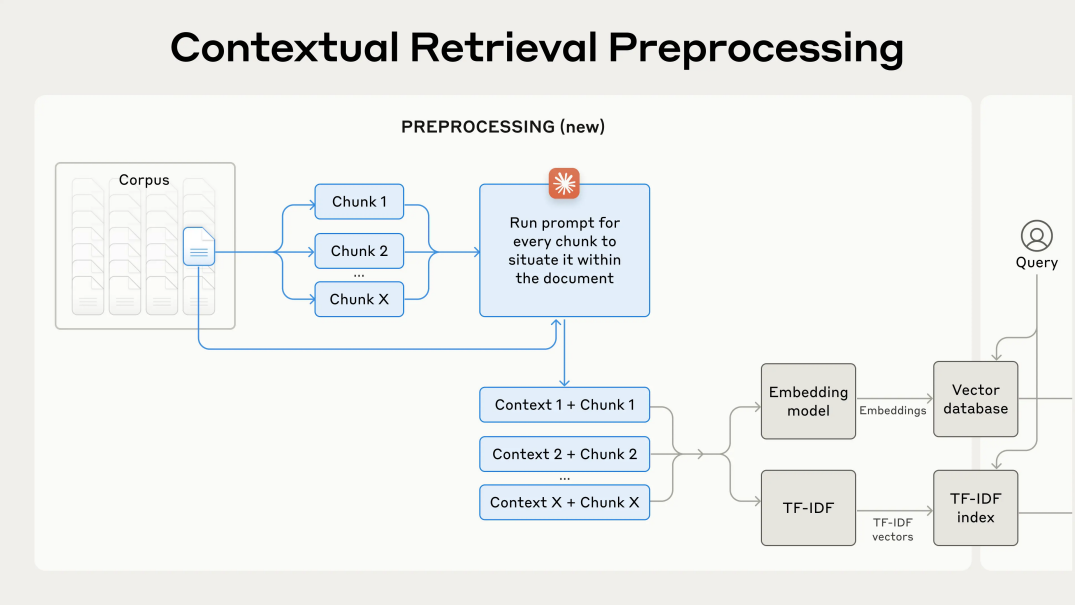

Figure 2: Contextual Preprocessing

3.2.3 Ensemble Retrieval with Reranking

The retrieval strategy leverages complementary strengths through a three-stage process. BM25 handles exact keyword matches and Vietnamese technical terms through custom tokenization with underthesea, while ChromaDB provides semantic similarity search using `dangvantuan/vietnamese-document-embedding` with the contextually enhanced content.

I implemented Cohere reranking (Cohere, 2017) as the final optimization stage, retrieving significantly more documents from the ensemble retriever ($n = top\_k * 5$) then applying Cohere's rerank-v3.5 model to identify the most relevant results. This approach ensures high recall from initial retrieval while achieving precision through neural reranking that understands semantic relationships beyond keyword matching.

The reranking process stores relevance scores in document metadata, enabling future analysis of retrieval quality and providing transparency for debugging. Each returned document includes both its original relevance from the ensemble retriever and its reranked score, supporting sophisticated result interpretation.

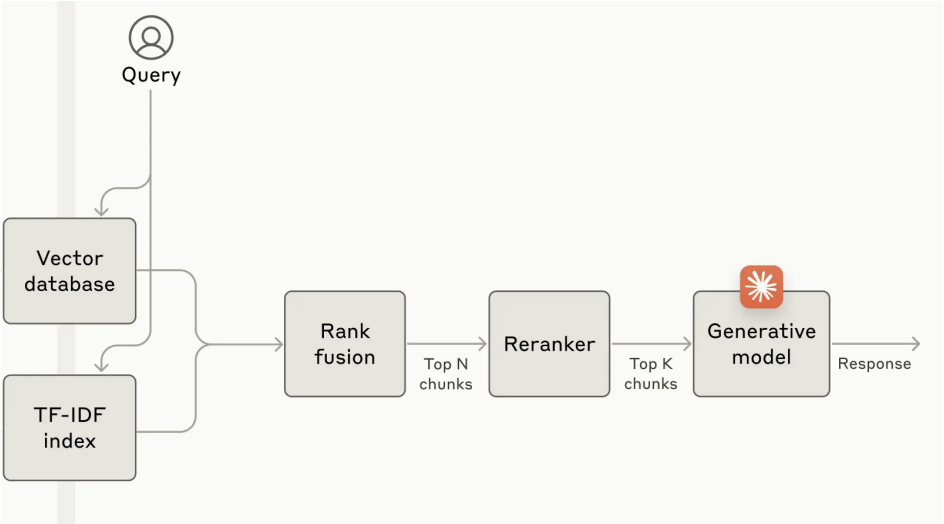

Figure 3: Ensemble Retrieval with Reranking

3.3 Multi-turn Conversation Management

3.3.1 Conversation State Architecture

I implemented short-term conversation memory through LangGraph's InMemorySaver with thread-based conversation tracking. Each conversation receives a unique thread_id that maintains state across multiple user interactions while enabling concurrent conversations without interference. The `ViettelPayState` TypedDict tracks comprehensive conversation metadata including message history, processing context, enhanced queries, and retrieved documents.

The `get_conversation_context` function extracts relevant conversation history using a configurable sliding window (default 6 messages) that balances context preservation with token efficiency. I format con-

versation history clearly for LLM consumption, distinguishing between user messages and AI responses while maintaining chronological order.

3.3.2 Query Enhancement Node

This specialized node analyzes complete conversation context to transform ambiguous follow-up questions into comprehensive search queries that capture the user's actual information needs.

The enhancement process performs three critical transformations. First, it analyzes overall conversation context by combining information from previous messages and understanding relationships between current questions and previously discussed topics. Second, it replaces unclear references ("nó", "thế", "vậy", "đó") with specific objects from conversation context, converting vague questions like "làm sao khắc phục?" into explicit queries like "làm sao khắc phục lỗi 606?". Third, it expands queries with related ViettelPay terminology and Vietnamese expression variants to improve retrieval comprehensiveness.

For example, a conversation progressing from "Mã lỗi 606 là gì?" to "Làm sao khắc phục?" becomes enhanced to "cách khắc phục lỗi 606 trên ViettelPay Pro", dramatically improving retrieval accuracy for the follow-up question.

3.3.3 Context-Aware Intent Classification

I enhanced intent classification to leverage conversation context, enabling accurate categorization of ambiguous follow-up questions. The classifier receives complete conversation history alongside the current message, allowing it to understand implicit intent based on conversation flow. Statements like "tôi vẫn chưa hiểu" receive appropriate classification based on previously discussed topics rather than defaulting to unclear intent.

3.3.4 Agent Orchestration

The LangGraph workflow implements a state machine with five primary nodes processing conversations through conditional routing. The enhanced workflow sequence follows: Intent Classification $\rightarrow$ Query Enhancement $\rightarrow$ Knowledge Retrieval $\rightarrow$ Response Generation, with conditional branching based on intent types.

Script-based intents (greeting, out_of_scope, human_request, unclear) bypass knowledge retrieval for immediate response generation using predefined templates from conversation scripts. Knowledge-based intents (faq, error_help, procedure_guide) follow the complete enhancement and retrieval pipeline to provide accurate, contextually relevant responses.

I implemented component binding through functools.partial, eliminating repeated initialization overhead by pre-binding LLM clients, knowledge retrievers, and text processors to their respective nodes. Error handling provides graceful degradation where node failures don't terminate conversations, with intent classification failures defaulting to clarification requests and retrieval failures prompting human escalation.

The workflow supports runtime parameter adjustment through configurable fields, enabling modification of retrieval parameters and generation settings without system restart. This flexibility enables performance tuning based on conversation patterns and user feedback while maintaining system stability

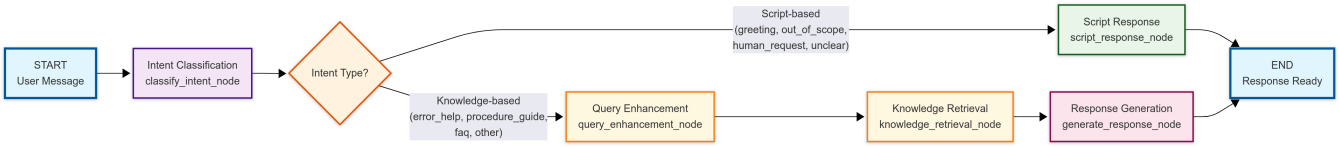

Figure 4: Lang Graph Workflow

4 Evaluation

4.1 Evaluation Framework Design

I designed a comprehensive evaluation framework to assess system performance across three critical dimensions: single-turn retrieval, multi-turn conversation handling, and intent classification accuracy. To ensure methodological rigor and prevent self-enhancement bias, I used GPT-4.1 for synthetic dataset generation while evaluating system performance with Gemini 2.0 Flash.

4.1.1 Dataset Generation Methodology

Single-Turn Retrieval Dataset: I processed all document chunks from the knowledge base to generate 2-3 questions per chunk using GPT-4.1. The generation prompt instructs the model to create diverse question types: direct information queries, procedural questions, error troubleshooting scenarios, and policy inquiries. Each question is designed to be answerable from its source chunk while using natural Vietnamese language patterns typical of ViettelPay Pro users. The dataset includes $\sim$ 300 question-document pairs covering error codes, transaction procedures, fee policies, and technical guidelines.

Multi-Turn Conversation Dataset: I generated $\sim$ 300 complex conversation scenarios that simulate realistic customer support interactions. GPT-4.1 creates 2-4 turn conversations with three scenario types: error resolution (users reporting problems and asking follow-up questions), procedure guidance (step-by-step assistance with progressive clarification), and policy information (general questions leading to specific cases). Each conversation includes natural transitions where users ask ambiguous follow-up questions like "làm sao khắc phục?" after discussing specific errors. The dataset contains conversations with varying complexity from simple clarifications to multi-topic discussions.

Intent Classification Dataset: I created $\sim$ 250 conversation scenarios that test intent recognition across realistic interaction patterns. Each conversation will have multiple turns and a ground truth label for them. The generation process emphasizes intent diversity with 70% topic-related messages and 30% natural conversational elements (greetings, unclear requests, escalation demands). GPT-4.1 generates both single-chunk scenarios (focused discussions) and multi-chunk scenarios (conversations spanning multiple topics).

Special attention is given to edge cases like ambiguous messages that depend on conversation context for proper classification, and critical intent scenarios that test error handling and human escalation detection.

4.1.2 Evaluation Metrics:

Mean Reciprocal Rank (MRR): Measures the average reciprocal rank of the first relevant result (Craswell, 2009). Higher values indicate relevant results appear earlier in rankings. latex

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{1}$$

Where $|Q|$ is the total number of queries and $rank_i$ is the rank position of the first relevant document for query i.

Hit@K: Percentage of queries where at least one relevant document appears in the top-K results. Hit@5 = 0.93 means 93% of queries find relevant information within the first 5 results. latex

$$Hit@K = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathbb{I}(rank_i \leq K) \tag{2}$$

Where $|Q|$ is the total number of queries, $rank_i$ is the rank position of the first relevant document for query i, K is the cutoff threshold, and $\mathbb{I}$ is the indicator function (1 if condition is true, 0 otherwise).

Accuracy: Percentage of correctly classified intents out of total predictions.

$$Accuracy = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{3}$$

Precision/Recall/F1: Multi-class classification metrics where precision measures prediction accuracy per class, recall measures coverage of actual cases per class, and F1 provides balanced performance assessment. Macro-averaging treats all classes equally regardless of support.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{4}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{5}$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{6}$$

$$Macro\text{-}Average = \frac{1}{C}\sum_{i=1}^{C} Metric_i \tag{7}$$

Where TP_i, FP_i, FN_i are True Positives, False Positives, and False Negatives for class i respectively, C is the total number of classes, and $Metric_i$ represents the metric value for class i.

4.1.3 Cross-LLM Evaluation Strategy

The evaluation methodology implements a cross-LLM approach where different models handle dataset creation and system evaluation. GPT-4.1 generates synthetic questions, conversations, and intent scenarios based on business documentation, while Gemini 2.0 Flash processes these inputs through the actual system workflow. This separation prevents the system from being optimized for its own training data distribution, ensuring more realistic performance assessment.

4.1.4 Data Limitations and Constraints:

The evaluation relies on synthetic data generated by GPT-4.1, which may not fully capture the complexity and variability of real user interactions with ViettelPay Pro. Synthetic data can miss nuanced user behavior patterns, informal language variations, and edge cases that emerge in production environments. However, synthetic generation enables controlled evaluation across diverse scenarios while ensuring comprehensive coverage of business documentation content.

Due to API cost constraints, I limited each evaluation dataset to approximately 300 samples. While this provides sufficient data for initial performance assessment, larger datasets would enable more robust statistical analysis and better generalization confidence. The dataset size represents a practical balance between evaluation thoroughness and resource limitations during development.

4.2 Single-Turn Retrieval Evaluation

4.2.1 Methodology

Single-turn retrieval evaluation measures the system's ability to find relevant information for isolated queries without conversation context. I evaluate using Mean Reciprocal Rank (MRR) and Hit@K rates, where Hit@K measures the percentage of queries where the source document appears in the top-K results, and MRR calculates the average reciprocal rank of the first relevant result.

The evaluation process generates questions from each document chunk, then searches the knowledge base using the same chunk as ground truth. This approach tests whether the contextual enhancement and ensemble retrieval successfully maintain the connection between synthetic questions and their source content.

4.2.2 Results

The single-turn retrieval evaluation demonstrates exceptional performance across all metrics, validating the effectiveness of contextual enhancement and ensemble retrieval design.

Metric	Score	Percentage
Hit@1	0.704	70.4%
Hit@3	0.881	88.1%
Hit@5	0.932	93.2%
Hit@10	0.980	98.0%
MRR	0.801	80.1%

Table 1: Single-Turn Retrieval Performance

The results in Table 1 show that 98% of questions find their source document in the top-10 results, with 93.2% achieving top-5 retrieval. The 80.1% MRR indicates that relevant documents typically appear in high-ranking positions, demonstrating both high recall and precision in the retrieval process.

4.3 Multi-Turn Retrieval Evaluation

4.3.1 Methodology

Multi-turn retrieval evaluation assesses the system's ability to maintain context across conversation turns and enhance ambiguous follow-up queries. I generate multi-turn conversations using GPT-4.1, then convert them to enhanced queries using the query enhancement node. The enhanced queries are evaluated against the same ground truth documents as their conversation sources.

This evaluation tests the complete multi-turn pipeline: conversation context extraction, query enhancement transformation, and retrieval performance with enhanced queries. The methodology measures both the query enhancement effectiveness and the retrieval system's ability to handle contextually enriched searches.

4.3.2 Results

Multi-turn retrieval maintains strong performance while showing expected degradation due to conversation complexity and query transformation challenges.

Metric	Score	Single-Turn Comparison
Hit@1	0.560	-14.4%
Hit@3	0.792	-8.9%
Hit@5	0.863	-6.9%
Hit@10	0.898	-8.2%
MRR	0.685	-11.6%

Table 2: Multi-Turn Retrieval Performance

The 16-point degradation from single-turn performance is expected for conversation complexity, yet 89.8% Hit@10 and 86.3% Hit@5 demonstrate that the query enhancement successfully preserves retrieval quality across conversation turns. The MRR of 0.685 shows that enhanced queries still achieve good ranking positions despite the additional complexity of context transformation.

4.4 Intent Classification Evaluation

4.4.1 Methodology

Intent classification evaluation measures the system's ability to categorize user messages across seven intent categories: greeting, faq, error_help, procedure_guide, human_request, out_of_scope, and unclear. I generate diverse conversation scenarios that include both straightforward intent examples and challenging edge cases with context dependencies.

The evaluation focuses on overall accuracy, per-class performance metrics, and critical intent handling for business-important categories like error_help and human_request. I also analyze performance differences between generation methods (single-chunk vs. multi-chunk scenarios) and agent flow types (script-based vs. knowledge-based intents).

4.4.2 Results

Intent classification achieves strong overall performance with particularly excellent results for critical business intents.

Metric	Score	Percentage
Overall Accuracy	0.804	80.4%
Macro Precision	0.779	77.9%
Macro Recall	0.787	78.7%
Macro F1	0.782	78.2%

Table 3: Intent Classification Overall Performance

Intent	Precision	Recall	F1 Score	Number of data
greeting	0.878	1.000	0.935	72
faq	0.859	0.766	0.810	175
error_help	0.845	0.878	0.861	180
procedure_guide	0.721	0.729	0.725	188
human_request	1.000	0.969	0.984	64
out_of_scope	1.000	1.000	1.000	20
unclear	0.590	0.605	0.598	119

Table 4: Per-Class Intent Performance

4.5 Generation Method and Flow Analysis

The evaluation framework also analyzes performance variations across different dataset generation methods and agent flow types, providing insights into system behavior under different complexity conditions.

Method	Accuracy	Macro F1	Messages	Complexity
single_chunk	0.810	0.716	620	Lower
multi_chunk	0.739	0.655	180	Higher

Table 5: Performance by Generation Method

Flow Type	Accuracy	Messages	Processing Path
script_based_flow	0.813	257	Direct response
knowledge_based_flow	0.801	543	Full retrieval pipeline

Table 6: Performance by Agent Flow Type

4.6 Performance Analysis

4.6.1 Retrieval System Effectiveness

The retrieval evaluation validates the contextual enhancement approach, with 98.0% Hit@10 for single-turn and 89.8% for multi-turn scenarios. The contextual embedding strategy successfully bridges semantic gaps between user queries and document content, particularly for Vietnamese financial terminology. The 0.116 MRR degradation in multi-turn scenarios reflects expected complexity while maintaining operationally acceptable performance.

4.6.2 Intent Classification Robustness

Intent classification demonstrates excellent performance for business-critical categories. The 87.8% recall for error_help ensures that system errors receive appropriate attention, while 96.9% recall for human_request guarantees effective escalation handling. The lower performance for unclear intent (59.8% recall) is acceptable since these messages inherently lack sufficient information for confident classification.

4.6.3 Multi-Turn Conversation Capability

The multi-turn evaluation confirms successful conversation context management and query enhancement. The system effectively transforms ambiguous follow-up questions into comprehensive search queries, maintaining retrieval quality across conversation turns. This capability enables natural conversational flows where users can ask progressive questions without repeating context.

4.6.4 Production Readiness Assessment

The evaluation results demonstrate production-ready performance across all critical dimensions. Single-turn retrieval exceeds 90% accuracy for practical scenarios (Hit@5), multi-turn handling maintains operational effectiveness, and intent classification reliably routes conversations to appropriate processing paths. The cross-LLM evaluation methodology provides confidence that performance will generalize to real user interactions beyond the training data distribution

5 Limitations and Future Enhancements

5.1 Current Limitations

Memory and Persistence: InMemorySaver provides only session-based memory without persistence across system restarts, limiting long-term user relationship building.

Scenario Coverage: The system handles reactive support only, lacking proactive engagement features like maintenance notifications or inactivity-based outreach.

Data Constraints: Evaluation relies on synthetic datasets ($\sim$ 300 samples each) that may not capture real user interaction complexity. No validation against actual ViettelPay Pro conversations.

External Dependencies: Critical functions depend on external APIs (OpenAI, Gemini, Cohere), creating potential availability and performance risks.

5.2 Future Enhancements

Persistent Memory: Implement database-backed conversation storage with user profiling for long-term context preservation and personalized assistance.

Proactive Features: Develop intelligent notifications, abandoned transaction recovery, and behavior-based service recommendations.

Real Data Integration: Establish privacy-compliant frameworks for collecting actual user conversations to refine system performance.

Advanced Analytics: Integrate conversation analytics for business intelligence, performance monitoring, and service optimization insights.

Multi-modal Support: Extend to voice interactions, document uploads, and image-based queries for comprehensive support scenarios.

6 Conclusion

I successfully developed a production-ready AI agent for Vietnamese customer support achieving 98.0% Hit@10 retrieval accuracy and 80.4% intent classification with excellent performance on critical business intents (87.8% error-help recall, 96.9% human-request recall).

The key technical contributions include adapting Anthropic's contextual retrieval for Vietnamese financial documents, implementing innovative multi-turn query enhancement, and establishing comprehensive cross-LLM evaluation methodology. The automated document processing pipeline eliminates manual overhead while preserving semantic relationships through LLM-generated context.

The system provides immediate business value through automated error resolution, consistent procedure guidance, and reliable escalation handling. Vietnamese language optimization and domain-specific knowledge integration create a scalable foundation for customer support automation in Vietnamese fintech applications.

This work demonstrates that careful application of modern retrieval-augmented generation techniques, combined with rigorous evaluation methodology, can create robust conversational AI systems for specialized domains and languages, establishing a template for similar applications beyond ViettelPay Pro.

7 Acknowledgments

I would like to express my sincere gratitude to my mentors, Do Khac Phong and Lam Xuan Thu, for their invaluable guidance and support throughout this project under the Viettel Digital Talent program. Their expertise and mentorship were instrumental in shaping both the technical direction and implementation approach of this AI agent system.

I also acknowledge Viettel Digital Talent program for providing the opportunity to work on this meaningful project that addresses real-world challenges in Vietnamese customer support automation. The program's emphasis on practical innovation enabled me to explore cutting-edge technologies while delivering tangible business value.

Special thanks to the open-source community and research teams at Anthropic, whose contextual retrieval methodology formed the foundation of this work, and to dangvantuan for developing the Vietnamese document embedding model that enabled effective semantic search for Vietnamese financial content.

References

Anthropic. (2024). *Introducing contextual retrieval.* Retrieved from https://www.anthropic.com/news/contextual-retrieval

Cohere. (2017). *Cohere reranker.* Retrieved from https://docs.cohere.com/v2/reference/rerank

Craswell, N. (2009). Mean reciprocal rank. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of database systems* (pp. 1703–1703). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-39940-9_488 doi: 10.1007/978-0-387-39940-9_488

dangvantuan. (2019). *Vietnamese embedding model.* Retrieved from https://huggingface.co/dangvantuan/vietnamese-embedding