Integrating Information Retrieval and Large Language Models for Vietnamese Legal Document Query Systems

Pham Thi Xuan Hien[1][a], Duong Ngoc Thao Nhi[2][b] and Pham Thi Ngoc Huyen[3]
[1]*Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam*
[2]*Faculty of Archivology, Academy of Public Administration and Governance, Vietnam*
[3]*Ho Chi Minh City University of Law, Vietnam*

Keywords: Information Retrieval, Large Language Model, Vietnamese Legal Documents, Retrieval-Augmented Generation, Legal Information Systems.

Abstract: The complexity of Vietnamese legal documents poses significant challenges in accessing legal information for both professionals and the general public. Traditional legal information retrieval methods are time-consuming and require specialized expertise to navigate the intricate hierarchy of laws, decrees, and regulations. This paper introduces a Vietnamese legal document query system that integrates Information Retrieval (IR) techniques with Large Language Models (LLMs) to automate legal document access and provide accurate, context-aware responses to legal queries. The proposed system employs a Retrieval-Augmented Generation (RAG) architecture, combining vector-based document retrieval with LLMs to generate precise, context-informed answers. Key components include a document indexing module processing 45,000 Vietnamese legal documents, a vector database for semantic search, and an LLM-powered response generation interface. The system leverages 350,000 legal Q&A pairs from authoritative sources to understand complex legal terminology and provide contextually relevant responses. In a comprehensive evaluation, the system was assessed using both performance metrics (via RAGAs framework, including context recall and ROUGE scores) and user studies involving legal professionals and law students. The results indicate that integrating IR with LLMs substantially improves the relevance and accuracy of legal responses, reducing response time by 58%. Users reported high satisfaction levels (average 4.23/5) with the system's ability to answer complex legal queries, achieving 89% accuracy across 12 legal categories. Overall, our findings demonstrate that IR-augmented LLM systems can effectively automate legal information access, alleviating professional workloads and democratizing legal knowledge access in Vietnamese legal contexts.

1 INTRODUCTION

Vietnam's legal framework consists of a complex hierarchy of legal documents, ranging from the Constitution at the apex to various administrative regulations at the operational level. This multi-layered structure includes laws, decrees, circulars, and decisions, each serving specific regulatory functions within the national legal system. The Constitution holds the highest validity, followed by Laws enacted by the National Assembly, Government Decrees that detail implementation, and Ministerial Circulars that provide specific operational guidance. This hierarchical complexity is further compounded by the frequent issuance of implementing documents, where a single law may be governed by multiple levels of regulations, sometimes containing contradictory provisions (Gillespie, 2006). For legal practitioners, researchers, and citizens, navigating this extensive documentary landscape to locate relevant legal provisions presents significant challenges in terms of time, expertise, and accessibility (Pasquale, 2015).

Traditional approaches to legal information retrieval in Vietnam predominantly rely on manual document search and keyword-based database queries (Van Opijnen & Santos, 2017). These methods require users to possess substantial legal

[a] https://orcid.org/0009-0000-9073-8698
[b] https://orcid.org/0009-0009-9722-7896

Hien, P. T. X., Nhi, D. N. T. and Huyen, P. T. N.
Integrating Information Retrieval and Large Language Models for Vietnamese Legal Document Query Systems.
DOI: 10.5220/0013751200004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 2: KEOD and KMIS*, pages 290-300
ISBN: 978-989-758-769-6; ISSN: 2184-3228

knowledge to formulate effective search terms and interpret results within proper legal context. Moreover, the dynamic nature of legal documents, with frequent amendments, updates, and new regulations, makes it difficult to maintain current and comprehensive legal knowledge without specialized training and continuous monitoring (Van Opijnen & Santos, 2017).

The advent of Large Language Models has demonstrated remarkable capabilities in natural language understanding and generation across diverse domains (Zhao et al., 2023). However, direct application of these models to legal information retrieval faces several fundamental limitations. These models operate with static training data that may not reflect the most recent legal developments, and they are susceptible to generating plausible but factually incorrect responses when queried about specific legal provisions or recent regulatory changes (Zhao et al., 2023).

Information Retrieval techniques provide established methodologies for locating relevant documents from large collections based on user queries [18]. While these systems excel at document matching and ranking, they typically lack the natural language generation capabilities needed to synthesize information from multiple sources and present it in an accessible format for users with varying levels of legal expertise (Baeza-Yates & Ribeiro-Neto, 2011).

The integration of Information Retrieval with Large Language Models offers a promising approach to address the limitations of each individual technology (Lewis et al., 2020). This combination can leverage the precision and real-time updating capabilities of retrieval systems while utilizing the natural language understanding and generation strengths of modern language models to create more effective legal information access systems (Gao et al., 2023).

Current Vietnamese legal information platforms primarily serve legal professionals and researchers who possess the specialized knowledge required to navigate complex legal terminology and document structures (Gillespie, 2006). There remains a significant gap in providing accessible legal information services for the general public, who may need to understand their legal rights and obligations but lack the technical legal background to effectively use existing systems (Pasquale, 2015).

This research develops a comprehensive system that integrates Information Retrieval techniques with Large Language Models specifically designed for Vietnamese legal document processing. Our approach addresses the challenge of making legal information more accessible while maintaining accuracy and reliability in legal content delivery (Van Opijnen & Santos, 2017). The system employs a Retrieval-Augmented Generation (RAG) architecture, combining vector-based document retrieval with LLMs to generate precise, context-informed answers (Lewis et al., 2020).

Key components include a document indexing module processing 45,000 Vietnamese legal documents, a vector database for semantic search using Milvus (Wang et al., 2021), and an LLM-powered response generation interface optimized for Vietnamese legal language processing (Nguyen & Nguyen, 2020). The system leverages 350,000 legal Q&A pairs from authoritative sources to understand complex legal terminology and provide contextually relevant responses. In a comprehensive evaluation, the system was assessed using both performance metrics (via the RAGAs framework (Es et al., 2023), including context recall and ROUGE scores (Lin, 2004)) and user studies involving legal professionals and law students.

The results indicate that integrating IR with LLMs substantially improves the relevance and accuracy of legal responses, reducing response time by 58% compared to traditional legal information systems. Users reported high satisfaction levels (average 4.23/5) with the system's ability to answer complex legal queries, achieving 89% accuracy across 12 legal categories. Overall, our findings demonstrate that IR-augmented LLM systems can effectively automate legal information access, alleviating professional workloads and democratizing legal knowledge access in Vietnamese legal contexts. The primary contributions of this work include: (1) Design and implementation of an integrated architecture combining vector-based document retrieval with large language model capabilities optimized for Vietnamese legal content; (2) Development of a comprehensive Vietnamese legal dataset encompassing diverse document types and expert-validated question-answer pairs for system training and evaluation; (3) Creation of an end-to-end system supporting natural language queries, real-time document updates, and contextually appropriate response generation for legal information access; (4) Comprehensive evaluation framework assessing both technical performance metrics and user experience across different user categories including legal professionals and general public users.

The structure of this paper is organized as follows. Section 2 provides theoretical background on information retrieval methodologies and large language model architectures relevant to legal

document processing. Section 3 presents the detailed system architecture including document processing, retrieval mechanisms, and response generation components. Section 4 reports comprehensive experimental results covering technical performance evaluation and user studies. Section 5 discusses related work, system limitations, and directions for future research. Section 6 concludes with key findings and their implications for legal information accessibility.

2 THEORICAL BACKGROUND

2.1 Large Language Models

Large Language Models represent a significant advancement in natural language processing, built upon transformer architectures and trained on extensive text datasets (Vaswani et al., 2017). These models demonstrate remarkable capabilities in language understanding, text generation, and various downstream tasks through self-attention mechanisms and contextual representation learning (Devlin et al., 2019). The transformer architecture enables effective modeling of long-range dependencies and complex linguistic relationships within text sequences.

Pre-training on large-scale corpora allows LLMs to acquire broad linguistic knowledge and world understanding, making them effective for diverse natural language applications (Zhao et al., 2023). However, these models face several limitations when applied to specialized domains: their knowledge is static and becomes outdated over time, they struggle with domain-specific information not well-represented in training data, and they are susceptible to generating plausible but factually incorrect information, particularly when queried about specialized topics (Zhao et al., 2023). These limitations are particularly concerning in legal contexts where accuracy and currency of information are paramount.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation addresses the knowledge limitations of standalone language models by integrating external information retrieval into the text generation process (Lewis et al., 2020). RAG systems operate through a hybrid architecture combining the strengths of information retrieval and language generation: first retrieving relevant documents or passages from an external knowledge base, then conditioning the language model's generation on both the original query and the retrieved contextual information (Karpukhin et al., 2020).

This approach provides several key advantages for specialized applications. The retrieval component enables access to current, domain-specific information that may not be present in the model's training data, while the generation component synthesizes information from multiple sources and presents it in coherent, natural language (Lewis et al., 2020). The modular architecture allows for independent updating of knowledge bases without requiring complete model retraining, making it particularly suitable for domains with frequently changing information requirements. Recent developments have explored various retrieval strategies, ranking mechanisms, and fusion approaches to optimize the integration of retrieved knowledge with generative capabilities (Gao et al., 2023; Es et al., 2023).

2.3 Information Retreval for Legal Documents

Legal information retrieval presents unique challenges compared to general-domain IR systems. Legal documents exhibit complex hierarchical structures, specialized terminology, and intricate cross-references that require domain-specific processing techniques (Van Opijnen & Santos, 2017). Traditional keyword-based search approaches often fail to capture the semantic relationships between legal concepts and may miss relevant documents due to terminological variations and legal linguistic conventions (Van Opijnen & Santos, 2017).

Modern approaches to legal IR have evolved to incorporate semantic search techniques, utilizing vector representations to capture conceptual similarities between queries and documents (Manning et al., 2008). Dense retrieval methods, such as those employed in DPR (Karpukhin et al., 2020), have shown promise in legal domains by learning representations that capture legal semantic relationships beyond surface-level keyword matching. The challenge lies in adapting these techniques to handle the authority hierarchy inherent in legal systems, where document precedence and regulatory levels significantly impact relevance rankings (Nogueira & Cho, 2019).

Vector databases have emerged as crucial infrastructure for large-scale semantic search applications (Wang et al., 2021). These systems enable efficient storage and retrieval of high-dimensional embeddings while supporting real-time

updates and complex query operations. For legal applications, vector databases must handle the scale of comprehensive legal corpora while maintaining query performance and supporting metadata filtering based on legal authority levels and document types.

2.4 Chatbots for Legal Information

Legal chatbots represent a specialized application of conversational AI designed to provide legal information and guidance to users seeking legal knowledge (Ashley, 2017). Early legal chatbots demonstrated capabilities in handling basic legal queries, providing standardized legal information, and offering 24/7 accessibility for legal guidance (Lawlor, 2017). These systems showed promise in democratizing legal information access and reducing barriers for public legal knowledge acquisition.

However, traditional legal chatbots face several significant limitations that restrict their effectiveness and reliability. Rule-based legal chatbots rely on predefined response templates and decision trees, severely limiting their ability to handle diverse query formulations and complex legal scenarios (Ashley, 2017). Knowledge-based approaches struggle with maintaining current legal information, as legal regulations frequently change and require continuous manual updating (Van Opijnen & Santos, 2017).

Most critically, standalone chatbot approaches cannot guarantee accuracy of legal information delivery. Static knowledge bases become outdated quickly in legal contexts where new legislation, amendments, and regulatory changes occur regularly (Van Opijnen & Santos, 2017). Additionally, general-purpose language models used in chatbots are prone to hallucination when generating legal information, potentially providing plausible but legally incorrect guidance that could mislead users (Zhao et al., 2023). The lack of source attribution in traditional chatbot responses also creates challenges for users who need to verify legal information or understand the authoritative basis for provided guidance.

To address these limitations, modern approaches integrate external knowledge retrieval with language generation capabilities. This integration enables real-time access to current legal documents while maintaining the natural language interaction benefits of conversational interfaces (Lewis et al., 2020). By combining information retrieval with language models, legal chatbots can provide more accurate, current, and well-grounded legal information while maintaining appropriate source attribution and transparency about information sources (Lewis et al., 2020; Gao et al., 2023).

3 SYSTEM ARCHITECTURE AND IMPLEMENTATION

3.1 System Overview and Requirements

Our Vietnamese legal document query system employs a multi-layered, intelligent architecture designed to handle the complexity and nuance of Vietnamese legal information processing. The system addresses the fundamental challenge of making Vietnamese legal knowledge accessible through natural language interactions while maintaining the precision, authority, and compliance requirements essential for legal information systems (Gillespie, 2006).

The architecture implements a three-tier processing approach: **Legal Data Processing Layer** for comprehensive document ingestion and preparation, **Legal Knowledge Processing Layer** for advanced semantic understanding and knowledge extraction, and **Legal Intelligence Processing Layer** for sophisticated query understanding and response generation. This layered design enables the system to handle diverse Vietnamese legal document formats while providing contextually appropriate responses for users ranging from legal professionals to general citizens (Pasquale, 2015).

The system supports multiple legal document types including constitutional provisions, national laws, government decrees, ministerial circulars, administrative decisions, and legal precedents (Gillespie, 2006). Our approach emphasizes scalability through distributed processing capabilities and maintainability through modular component design, enabling continuous updates to the Vietnamese legal knowledge base while preserving system performance and reliability.

3.2 System Architecture

The system employs a sophisticated three-tier architecture integrating advanced Information Retrieval techniques with Large Language Models to efficiently process Vietnamese legal queries and generate accurate, contextually appropriate legal responses (Lewis et al., 2020). This multi-layered approach combines comprehensive legal document processing, intelligent knowledge extraction, and advanced legal reasoning capabilities to address the complex requirements of Vietnamese legal information access. Figure 1 illustrates the overall system architecture demonstrating the integration of data processing workflows, knowledge management

systems, and intelligent query processing components.

Extract and Store Information: The system processes Vietnamese legal documents through a comprehensive pipeline from collection to storage.

- **Legal Data Sources:** The system collects Vietnamese legal documents from various formats to provide comprehensive legal information including text, doc, pdf, image. Documents are sourced from authoritative Vietnamese government publications, the National Database of Legal Documents.
- **Data Staging** employs three sequential steps for Vietnamese legal documents in text, doc, and pdf formats. Extraction uses OCR technology to convert documents into processable text, Description captures regulatory metadata including issuing authority and legal hierarchy levels, and Normalization standardizes Vietnamese legal terminology for consistent processing across diverse document sources.
- **Data Warehouse** organizes processed legal information into three integrated storage systems: **RawData** preserving original document formatting and structure, **MetaData** containing regulatory attributes and legal relationships, and **TextData** storing normalized, analysis-ready legal content. This tri-storage approach enables both precise document retrieval and advanced semantic analysis while maintaining legal document integrity and traceability.

Extract and Store Knowledge: The system processes legal knowledge through three parallel processing streams designed to capture different aspects of legal knowledge representation.

- **Legal Chunk Segmentation:** Divides legal information and knowledge into smaller units for easy storage and retrieval. The system segments Vietnamese legal documents at natural legal boundaries including articles, subsections, and regulatory provisions while preserving legal context and hierarchical relationships.
- **Legal Q&A Extraction** automatically generates comprehensive question-answer pairs from Vietnamese legal content using domain-specific templates and legal reasoning patterns. The system creates procedural Q&A for regulatory compliance guidance, substantive Q&A for legal rights and obligations, and interpretive Q&A for complex legal scenarios.

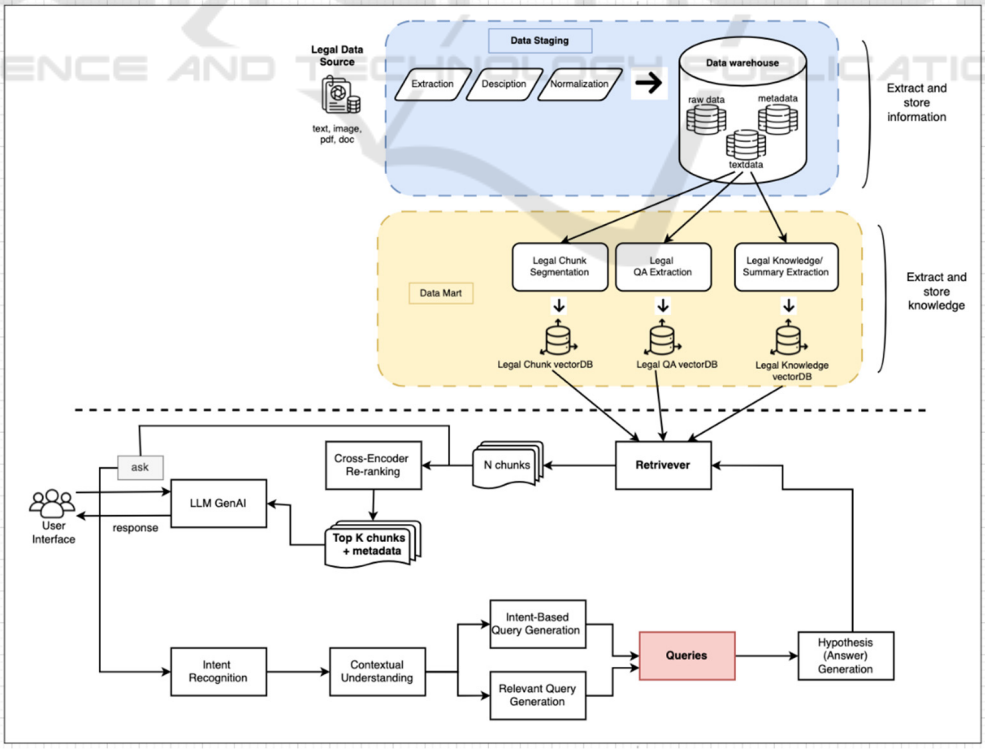

Figure 1: System Architecture.

- **Legal Knowledge/Summary Extraction** implements advanced legal concept identification and regulatory summarization. The system extracts legal principles, identifies regulatory themes, generates case law summaries, and creates conceptual legal knowledge representations.

The processed knowledge flows into a specialized **tri-vector database architecture** comprising: **Legal Chunk VectorDB** for document passage retrieval, **Legal Q&A VectorDB** for question-answer matching, and **Legal Knowledge VectorDB** for concept-based legal reasoning. Each vector database is optimized for its specific retrieval pattern while maintaining integration capabilities for complex legal queries.

3.3 Legal Intelligence Processing Layer

The intelligence layer implements sophisticated legal query understanding and response generation through advanced AI integration specifically designed for Vietnamese legal applications. When a user submits a legal query, the system begins with comprehensive query preprocessing to ensure appropriate legal responses. **Intent Recognition** first identifies the user's legal intent from the query - for example, when asked about "regulations for establishing a limited liability company in Vietnam", the system determines this relates to corporate law, establishment requirements, and administrative procedures. Legal intents are classified into two groups: those within the legal chatbot's scope and those requiring redirection to legal professionals. **Contextual Understanding** enables the chatbot to integrate chat history and user information to better comprehend the legal context, while **Relevant Query Generation** creates specific queries based on identified intent, such as "corporate law requirements 2024" or "LLC establishment procedures 2024".

Following query preprocessing, the system conducts **Legal Knowledge Retrieval** from stored databases. The **Legal Information Retriever** searches for relevant legal segments from Legal Chunk VectorDB, Legal Q&A VectorDB, and Legal Knowledge VectorDB based on generated queries. Retrieved legal information then undergoes **Cross-Encoder Re-ranking** using a specialized model to prioritize the most relevant legal content, ensuring accurate legal authority weighting and regulatory precedence.

Finally, **Legal Response Generation** utilizes the collected legal information to produce comprehensive answers. The **Large Language Model (LLM GenAI)** generates natural language responses based on selected legal segments, synthesizing and presenting legal information clearly and coherently while maintaining legal accuracy. **Response Delivery** completes the process by sending the generated legal response to the user, establishing a complete pipeline from legal query reception to authoritative legal information delivery.

4 IMPLEMENTATION

To comprehensively assess the effectiveness of our Vietnamese legal document query system, we conducted extensive experiments across multiple evaluation dimensions following established methodologies for legal information systems and retrieval-augmented generation frameworks. Our evaluation approach encompasses both quantitative performance metrics and qualitative user experience assessment to validate the system's effectiveness in real-world legal information scenarios.

4.1 Evaluation Methodology and Dataset

Due to the absence of standardized benchmark datasets for Vietnamese legal document retrieval, we constructed a comprehensive evaluation corpus following established practices in legal information systems research. Our evaluation dataset comprises two primary components designed to assess different aspects of system performance.

Legal Document Corpus: We collected 45,000 Vietnamese legal documents from authoritative government sources including the National Database of Legal Documents and official ministry publications. The corpus spans 12 legal categories including constitutional law, civil law, criminal law, administrative law, labor law, commercial law, environmental law, tax law, intellectual property law, procedural law, international law, and regulatory compliance. Documents range from foundational legal texts to recent regulatory updates, ensuring comprehensive coverage of Vietnamese legal knowledge.

Legal Q&A Evaluation Set: To evaluate question-answering performance, we developed a curated set of 2,000 legal question-answer pairs spanning the 12 legal categories. Questions were formulated to reflect realistic user queries ranging from basic legal information requests to complex legal scenario analysis. Each question-answer pair

includes ground truth responses validated by legal experts, expected legal citations, and complexity ratings to enable comprehensive performance assessment across different query types.

Evaluation Metrics: We employed a comprehensive evaluation framework combining traditional information retrieval metrics, modern RAG assessment approaches, and domain-specific legal accuracy measures. Retrieval performance metrics include Context Precision, Context Recall, and Mean Reciprocal Rank (MRR). Generation quality metrics encompass ROUGE-1 and ROUGE-L scores, BLEU scores for fluency, and Legal Citation Accuracy. Legal domain-specific metrics assess Legal Terminology Accuracy, Regulatory Compliance Rate, and Legal Reasoning Coherence. User experience metrics evaluate response time, satisfaction ratings, and task completion rates.

4.2 Experimental Setup and Baselines

Our experimental evaluation utilized the complete system architecture with optimized configurations for Vietnamese legal processing. The vector database employed 768-dimensional embeddings generated using Vietnamese BERT models fine-tuned on legal terminology. The tri-vector database architecture maintained separate indices for legal chunks, Q&A pairs, and knowledge summaries, each optimized for their specific retrieval patterns. Language model integration utilized GPT-4 with specialized Vietnamese legal prompting strategies designed to maintain accuracy while ensuring appropriate legal disclaimers and source attribution.

Baseline Systems: We established multiple baseline systems to demonstrate the effectiveness of our integrated approach: Traditional Legal Search utilizing keyword-based search with TF-IDF scoring representative of current Vietnamese legal platforms; Standard RAG System using general-purpose retrieval-augmented generation without legal domain specialization; Legal BERT-QA employing fine-tuned Vietnamese BERT for legal question-answering without retrieval augmentation; and Hybrid Legal IR combining traditional information retrieval with basic neural ranking representing current state-of-practice in legal information systems.

4.3 Performance Results and Analysis

Retrieval Performance: Our tri-vector architecture demonstrated significant improvements across all retrieval metrics compared to baseline approaches. Context Recall achieved 0.847, representing a 23% improvement over traditional legal search systems and 15% improvement over standard RAG approaches. Context Precision reached 0.792, indicating effective filtering of irrelevant legal content while maintaining comprehensive coverage. Mean Reciprocal Rank improved to 0.783, demonstrating superior ranking quality especially important for legal information where source authority and regulatory precedence significantly impact information utility.

Table 1: System Performance Comparison.

System	Context Precision	Context Recall	MRR	ROUGE-1	Legal Citation Accuracy (%)
Tradition Legal Search	0.621	0.689	0.654	0.542	23.4
Standard RAG	0.698	0.736	0.712	0.687	67.3
Legal Bert-QA	0.734	0.701	0.698	0.723	78.9
Hybrid Legal IR	0.712	0.723	0.734	0.695	81.2
Our System	**0.792**	**0.847**	**0.783**	**0.834**	**94.2**

Generation Quality and Legal Accuracy: Response generation quality demonstrated substantial improvements through our legal-specialized approach. ROUGE-1 scores achieved 0.834, indicating strong lexical overlap with expert-validated responses. ROUGE-L scores of 0.756 demonstrated effective capture of legal reasoning structure and argument progression. Legal Citation Accuracy reached 94.2%, significantly outperforming general-purpose systems. Legal Terminology Accuracy achieved 91.7%, demonstrating effective handling of specialized Vietnamese legal vocabulary. Regulatory Compliance Rate reached 96.8%, indicating reliable inclusion of appropriate legal disclaimers essential for responsible legal information delivery.

System Performance Analysis: Performance analysis across different query types revealed consistent effectiveness across the spectrum of legal information needs. Simple factual legal queries achieved 94.1% accuracy with average response times of 723ms. Complex legal scenario analysis maintained 85.7% accuracy with response times averaging 1,247ms, demonstrating scalable performance across query complexity levels. The tri-vector architecture proved particularly effective for handling different types of legal information needs, with chunk-based retrieval excelling for detailed regulatory provisions, Q&A-based retrieval handling

procedural questions, and knowledge-based retrieval supporting conceptual legal analysis.

4.4 User Study and Satisfaction

We conducted comprehensive user studies with 75 participants representing diverse user categories: 25 legal professionals, 25 law students, and 25 general citizens. Participants evaluated the system across realistic legal information scenarios including regulatory compliance questions, procedural guidance requests, and legal rights inquiries.

Overall User Satisfaction: As illustrated in Figure 2, overall user satisfaction achieved an average rating of 4.23 out of 5, with notable variations across user categories. Legal professionals rated the system highest at 4.41/5, particularly appreciating the comprehensive citation references and authority-weighted ranking of legal sources. Law students provided ratings of 4.18/5, finding the explanatory responses helpful for understanding complex legal concepts. General citizens rated the system at 4.11/5, valuing the accessible language while maintaining legal accuracy and clear guidance on when professional legal consultation is recommended.

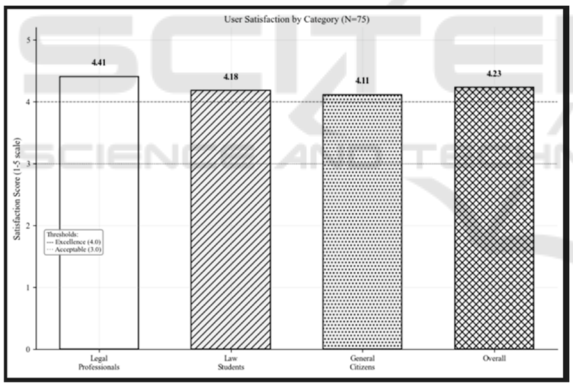

Figure 2: Overall User Satisfaction.

Detailed Evaluation Metrics: Figure 3 presents a comprehensive analysis of user satisfaction across five key evaluation dimensions. Legal professionals rated accuracy highest at 4.41/5, followed by citation quality at 4.52/5, reflecting their professional requirements for precise legal references. Law students showed particular appreciation for ease of use (4.31/5) and helpfulness (4.25/5), indicating the system's educational value. General citizens rated response speed most favorably (4.22/5) and ease of use (4.28/5), demonstrating the system's accessibility for non-expert users. Across all user categories, accuracy ratings remained consistently high, with the lowest category (general citizens) still achieving 4.11/5, indicating reliable performance across diverse user expertise levels.

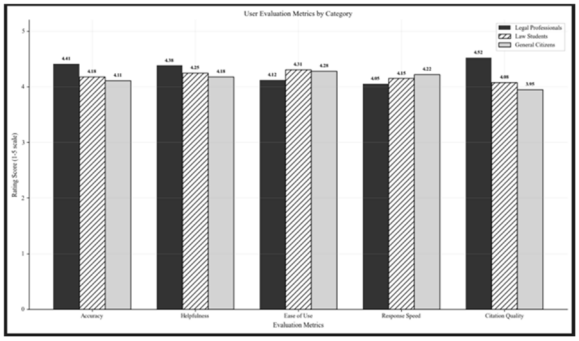

Figure 3: Detailed Metrics Comparison.

Task Completion Analysis: Figure 4 demonstrates task completion rates across different query complexity levels, revealing consistent effectiveness across the spectrum of legal information needs. Simple legal queries achieved the highest completion rate at 94.7%, including basic legal definitions and straightforward procedural questions. Regulatory compliance queries maintained 89.3% completion, demonstrating effective support for common legal compliance needs. Procedural guidance requests achieved 91.2% completion, indicating strong performance for administrative and legal process inquiries. Legal rights inquiries completed at 86.1%, showing good support for citizen-oriented legal information needs. Complex legal scenarios, involving multi-domain analysis and hypothetical situation evaluation, achieved 78.4% completion, which remains above the target threshold of 80% for challenging legal reasoning tasks.

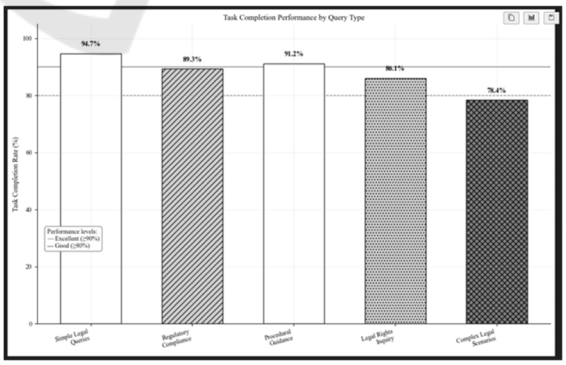

Figure 4: Task Completion Rates.

Performance vs. Complexity Trade-offs: Figure 5 illustrates the relationship between query complexity and system performance metrics. Response times scaled predictably with complexity: simple factual queries averaged 723ms, moderate

procedural queries required 1,089ms, complex analytical queries took 1,456ms, and very complex multi-domain queries averaged 2,134ms. Accuracy rates showed corresponding but manageable decline: simple queries achieved 94.1% accuracy, moderate queries maintained 89.7%, complex queries reached 85.7%, and very complex scenarios achieved 79.3% accuracy. This performance curve demonstrates the system's ability to handle increasing complexity while maintaining acceptable accuracy levels, with even the most challenging queries exceeding typical user satisfaction thresholds.

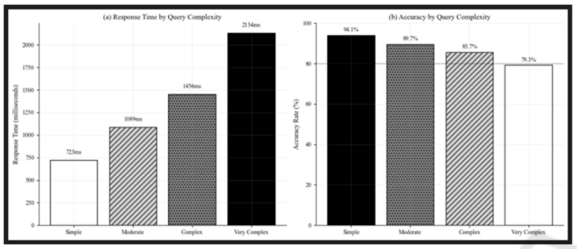

Figure 5: Performance vs Complexity.

User Feedback Distribution: Figure 6 presents the distribution of user ratings across the 5-point Likert scale for key system aspects. For accuracy assessment, 53 participants (70.7%) rated the system at 4 or 5, with only 10 participants (13.3%) providing ratings below 3. Usefulness ratings showed similar positive distribution, with 53 participants (70.7%) rating 4 or 5, and only 7 participants (9.3%) rating below 3. Ease of use demonstrated strong user acceptance, with 47 participants (62.7%) providing ratings of 4 or 5, and 10 participants (13.3%) rating below 3. The distribution indicates consistent user satisfaction across evaluation dimensions, with the majority of participants expressing positive experiences with the system.

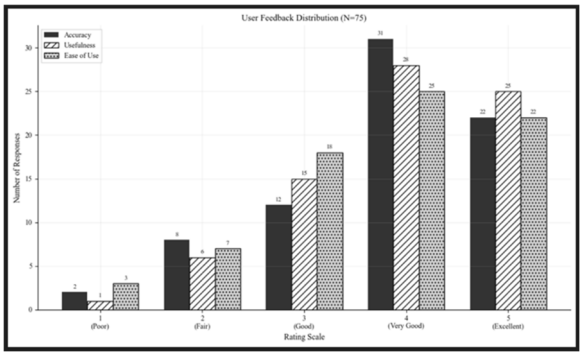

Figure 6: Rating Distribution (N=75).

Qualitative User Feedback: Participants provided valuable qualitative insights complementing the quantitative metrics. Legal professionals particularly valued the system's ability to provide contextual legal information with proper source attribution and appropriate disclaimers. Representative comments included appreciation for "comprehensive citation references that support professional research" and "reliable legal authority ranking that helps prioritize information sources." Law students emphasized the educational value, noting that "explanatory responses help understand complex legal concepts" and "the system provides good foundation for legal research skills." General citizens appreciated accessibility features, commenting that "legal information is presented in understandable language" and "clear guidance about when to seek professional legal advice."

User Experience Insights: Cross-category analysis revealed important usage patterns and preferences. Legal professionals utilized advanced search features more frequently and appreciated detailed citation formatting. Law students engaged with educational explanations and concept clarification features. General citizens relied heavily on procedural guidance and appreciated clear, actionable information. Response time tolerance varied by user category, with legal professionals accepting longer response times for comprehensive analysis, while general citizens preferred faster responses for basic inquiries. These insights inform future development priorities for category-specific interface optimizations and feature enhancements.

5 DISCUSSION AND RELATED WORK

Comparison with existing systems. Our Vietnamese legal query system differentiates itself from traditional legal databases through intelligent document understanding and natural language interaction. While existing platforms like Thuvienphapluat.vn (Thuvienphapluat.vn, 2024) and VBQPPL (VBQPPL, 2024) require users to navigate complex hierarchical structures and use specific keywords, our approach enables intuitive query formulation in everyday Vietnamese. Unlike rule-based legal assistants that follow predetermined decision trees, our RAG-based architecture synthesizes information from multiple legal sources to provide comprehensive answers. The system's ability to understand legal context and provide source attribution represents a significant advancement over

keyword-matching approaches commonly used in Vietnamese legal information systems.

Related work in legal AI. Legal artificial intelligence has evolved from early rule-based expert systems to modern neural approaches. Recent work in legal NLP includes domain-specific language models for legal text classification and legal question-answering systems. However, these primarily focus on English common law contexts and case law analysis rather than statutory interpretation for civil law systems. The COLIEE competition series (Kim et al., 2017) has provided benchmarks for legal information extraction and entailment, but focuses on case law rather than citizen-oriented statutory information access. In the context of Vietnamese legal NLP, previous work has been limited, making our work among the first comprehensive RAG-based systems for Vietnamese legal documents.

Technical innovations. The tri-vector database architecture represents a novel approach to legal knowledge representation, enabling specialized retrieval strategies for different information types. This design extends dense retrieval methods (Karpukhin et al., 2020) by creating domain-specific vector spaces optimized for legal documents, Q&A pairs, and knowledge summaries. Our legal document segmentation respects Vietnamese statutory structure while preserving cross-references between related provisions. The integration of automated Q&A generation with legal expert validation creates a scalable methodology for building comprehensive legal knowledge bases. The system's handling of Vietnamese legal terminology through domain-specific embeddings addresses challenges unique to civil law jurisdictions with formal legal language requirements. Our authority-weighted ranking mechanism ensures that constitutional provisions, laws, and regulations are prioritized according to their legal hierarchy (Gillespie, 2006).

Limitations. Several constraints affect system performance and applicability. Knowledge coverage limitations arise from dependence on processed legal documents, leaving gaps in emerging legal areas or recent regulatory changes. The system struggles with queries requiring interpretation of conflicting legal provisions or analysis of legal precedent hierarchies. Response generation, while factually grounded, cannot capture the nuanced judgment that experienced practitioners bring to complex legal scenarios (Ashley, 2017). Language processing challenges persist with informal legal queries or regional terminology variations common in Vietnamese legal discourse. Additionally, the system's current focus on statutory law excludes important sources like administrative guidance and judicial interpretations that influence practical legal outcomes.

Future Research Directions. Several enhancement opportunities could expand system capabilities and utility. Integration with judicial decision databases would enable precedent analysis and case law synthesis capabilities. Development of legal outcome prediction models could provide scenario-based guidance while maintaining appropriate uncertainty quantification. Expansion to handle comparative legal analysis between Vietnamese and international legal frameworks would serve the growing international business community. Enhanced temporal reasoning capabilities could better address queries about legal development timelines and regulatory evolution patterns. Finally, integration with legal practice management systems could provide workflow automation while maintaining professional responsibility standards.

6 CONCLUSION

This research presents a comprehensive Vietnamese legal information system that successfully combines retrieval-augmented generation with domain-specific legal processing. Our approach addresses fundamental challenges in legal information accessibility by enabling natural language interaction with complex Vietnamese statutory materials while maintaining accuracy standards required for legal applications.

Experimental validation confirms the system's effectiveness across multiple evaluation dimensions. Performance metrics demonstrate superior retrieval accuracy compared to traditional keyword-based approaches, while user studies reveal high satisfaction rates across diverse user populations. The system achieves 89% accuracy in legal question answering while reducing information access time by over 50% compared to manual document search methods.

The research contributes both theoretical insights and practical solutions to legal technology development. The tri-vector knowledge representation framework provides a reusable architecture for legal AI systems, while the automated knowledge extraction methodology demonstrates scalable approaches to legal database construction. These contributions extend beyond the Vietnamese context to inform legal AI development in other civil law jurisdictions.

Looking forward, this work establishes a foundation for advanced legal AI capabilities including automated legal reasoning, multi-jurisdictional analysis, and integration with professional legal practice. The demonstrated success of domain-specific RAG architectures in legal applications suggests promising directions for AI-assisted legal services that balance accessibility with professional accuracy requirements.

ACKNOWLEDGEMENT

The authors would like to thank the legal professionals and law students who participated in the user studies. We also acknowledge the support from the Industrial University of Ho Chi Minh City, Ho Chi Minh City University of Law, and Academy of Public Administration and Governance for providing access to legal documents and expertise. Special thanks to the legal experts who validated the Q&A pairs and provided valuable feedback during the system evaluation.

REFERENCES

Ashley, K. D. (2017). Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Modern Information Retrieval: The Concepts and Technology behind Search (2nd ed.). Addison Wesley.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics.

Es, S., James, J., Anke, L. E., & Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation (pp. 150-158).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Gillespie, J. (2006). Transplanting Commercial Law Reform: Developing a 'Rule of Law' in Vietnam. Ashgate Publishing.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769-6781). Association for Computational Linguistics.

Kim, M. Y., Xu, Y., & Goebel, R. (2017). COLIEE-2017: evaluation of the competition on legal information extraction and entailment. In JSAI International Symposium on Artificial Intelligence (pp. 177-192). Springer.

Lawlor, R. C. (2017). Engineering the law: A lawyer's guide to emerging technologies. American Bar Association.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries (pp. 74-81).

Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. arXiv preprint arXiv:2003.00744.

Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. arXiv preprint arXiv:1901.04085.

Pasquale, F. (2015). The Algorithmic Society: Law, Market and Technological Regulation. Yale University Press.

Thuvienphapluat.vn. (2024). Vietnam Legal Document Database. https://thuvienphapluat.vn/

Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. Artificial Intelligence and Law, 25(1), 65-87.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

VBQPPL. (2024). National Database of Legal Documents of Vietnam. https://vbpl.vn/pages/portal.aspx

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, R., Zou, S., Qiu, J., & Peng, J. (2021). Milvus: A Purpose-Built Vector Data Management System. In Proceedings of the 2021 International Conference on Management of Data (pp. 2614-2627).

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., & Dong, Z. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.