JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2026**, 26, 076-081

GIẢM THIỂU SỰ CHÊNH LỆCH VỀ TRÌNH ĐỘ HIỂU BIẾT PHÁP LÝ: HỆ THỐNG DỰA TRÊN RAG ĐỂ ĐIỀU HƯỚNG KHUÔN KHỔ PHÁP LÝ VIỆT NAM

Phan Van Nam [1*], Nguyễn Trung Hiếu[2]
[1] *Trường Đại học Nông Lâm, Phường Linh Xuân, Thành phố Hồ Chí Minh, Việt Nam*
[2] *Trường Đại học Lạc Hồng, Số 10 Đường Huỳnh văn Nghệ, Phường Trấn Biên, Tỉnh Đồng Nai, Việt Nam*
* Tác giả liên hệ: *phanvannambd@gmail.com*

THÔNG TIN BÀI BÁO

Ngày nhận: 05/11/2025
Ngày hoàn thiện: 02/03/2026
Ngày chấp nhận: 02/03/2026
Ngày đăng: 15/03/2026

TỪ KHÓA

Large Language Models;
Retrieval-augmented Generation;
Artificial intelligence;
Policymakers
Legal context.

TÓM TẮT

Nghiên cứu này đánh giá tiềm năng của mô hình ngôn ngữ lớn (Large Language Models - LLM) như một giải pháp hiệu quả về chi phí nhằm thu hẹp khoảng cách tiếp cận pháp lý tại Việt Nam. Bài viết phân tích tính khả thi của việc triển khai mô hình GPT tăng cường truy xuất (RAG), xem xét các thách thức về hạ tầng, khung pháp lý, độ chính xác của trí tuệ nhân tạo và khung quy định pháp lý địa phương. Từ các nghiên cứu thực tiễn quốc tế, nghiên cứu đề xuất các khuyến nghị chính sách nhằm tích hợp trí tuệ nhân tạo một cách có trách nhiệm, đảm bảo tính công bằng và quyền riêng tư.

MITIGATING DISPARITIES IN LEGAL LITERACY: A RAG-BASED SYSTEM FOR NAVIGATING THE VIETNAMESE LEGAL FRAMEWORK

Phan Van Nam [1*], Nguyễn Trung Hiếu
[1] *Nong Lam University, Linh Xuan Ward, Ho Chi Minh City, Vietnam*
[2]*Lac Hong University, No.10, Huynh Van Nghe Str,Tran Bien Ward, Dong Nai Province, Vietnam*
*Corresponding Author: *phanvannambd@gmail.com*

ARTICLE INFO

Received: Nov 5th, 2025
Revised: Mar 2nd, 2026
Accepted: Mar 2nd, 2026
Published: Mar 15th, 2026

KEYWORDS

Large Language Models;
Retrieval-augmented Generation;
Artificial intelligence;
Policymakers;
Legal context.

ABSTRACT

Limited legal access for marginalized populations in Vietnam creates significant disparities in justice. This paper investigates the potential of Large Language Models (LLMs) as a cost-effective solution to provide preliminary legal guidance and document preparation for underserved communities. We examine the feasibility of deploying a retrieval-augmented Generation (RAG) Generative Pre-training Transformer (GPT) framework, analyzing the requisite technological infrastructure, legal frameworks, and challenges such as Artificial intelligence (AI) accuracy and the interpretation of local legal nuances. Drawing from international case studies, this research offers recommendations for policymakers to integrate AI responsibly, ensuring fairness, privacy, and accessibility in legal service delivery.

Doi: https://doi.org/10.61591/jslhu.26.1019
Available online at: https://lhj.vn

1. INTRODUCTION

Many nations confront substantial challenges in enhancing legal literacy, particularly among marginalized populations. Globally, persistent barriers stemming from cultural, linguistic, and structural factors impede the effective engagement of these communities with national legal systems. This deficit often exacerbates inequality and curtails the effective administration of justice. These systemic challenges are particularly pronounced in Southeast Asia, and this paper focuses on the specific case of Vietnam. In Vietnam, significant barriers to justice for marginalized groups are well-documented, especially concerning geographic and linguistic isolation [1, 2]. For many of the nation's ethnic minority groups, Vietnamese is not their primary language.

This creates a significant linguistic divide that obstructs their comprehension of legal documents and procedures, exacerbating their vulnerability. Geographic isolation further compounds this challenge, as legal services and educational resources are predominantly concentrated in urban centers, leaving rural populations markedly underserved. Moreover, women in rural areas encounter unique obstacles rooted in entrenched cultural norms and systemic gender disparities, which restrict their access to legal information [3].

Concurrently, many ethnic minority communities traditionally rely on informal dispute resolution mechanisms. While culturally significant, this reliance often disengages them from the formal legal system, limiting their understanding of national laws and their attendant rights. Existing state-led legal literacy programs in Vietnam are frequently constrained by inadequate funding, outdated pedagogical methodologies, and a deficit of localized, culturally sensitive content. Furthermore, legal educators, such as local officials and legal aid workers, often lack the specialized training required to maximize the impact of these initiatives. These deficiencies underscore an urgent need for more inclusive, accessible, and sufficiently resourced legal literacy efforts [4, 5].

In response to the challenges identified in the Vietnamese context, this paper proposes and evaluates a novel technological intervention: a legal assistance tool integrating Retrieval-Augmented Generation (RAG) with the GPT LLM. This approach is posited as a promising solution to enhance legal accessibility for marginalized communities. The RAG architecture, which synergizes the parametric memory (pre-trained knowledge) of the language model with a non-parametric memory (retrieved external data), is particularly suitable. This method grounds the model's outputs in authoritative legal sources, thereby ensuring accurate, context-specific, and culturally sensitive responses while critically mitigating the risk of model "hallucinations".

In response to such challenges, the concept of RAG has gained significant attention, particularly in fields where contextual accuracy and domain-specificity are critical. RAG combines generative models, such as GPT-4, with retrieval mechanisms that leverage external databases to produce responses that are both contextually relevant and factually grounded. This hybrid approach directly addresses the limitations of standalone generative models, such as "hallucinations" or inaccuracies when dealing with domain-specific knowledge.

RAG frameworks integrate parametric memory (pre-trained knowledge within the language model) with non-parametric memory (external, retrievable data sources) to provide grounded outputs. The authors in [6, 7] introduced RAG as a method to enhance knowledge-intensive tasks, noting that its ability to dynamically retrieve relevant information ensures higher accuracy compared to static generative models. This capability is particularly advantageous in the legal domain, where responses must align with existing legal codes, case law, and procedural rules.

In the legal context, RAG's architecture allows for the retrieval of up-to-date legal texts, such as statutes, regulations, and case precedents, thereby reducing reliance on outdated or incomplete training data. Hybrid retrieval methods, combining keyword-based approaches like BM25 with semantic search using dense embeddings, have been shown to improve both precision and recall in information retrieval tasks [8]. These mechanisms ensure that outputs are not only relevant but also sufficiently comprehensive to address complex legal queries.

AI-powered legal systems leveraging RAG have been deployed in various jurisdictions to support tasks such as legal research, document drafting, and preliminary case analysis. In the United States, tools like Casetext's use retrieval-based approaches to assist attorneys in drafting legal documents [9, 10]. Similarly, in Europe, initiatives such as the LEXIA project have explored integrating retrieval-augmented models to automate routine legal tasks, focusing on improving efficiency and reducing costs [11]. These applications demonstrate RAG's potential to enhance legal accessibility, particularly for non-specialist users.

While the benefits of RAG systems are well-documented, their application is not without challenges. Legal text is characterized by high complexity, dense syntax, specialized terminology, and jurisdiction-specific nuances. Achieving consistent accuracy in legal reasoning tasks requires significant customization of both the retrieval and generation components [12]. Ethical considerations are also critical. Biases in training data can propagate through AI systems and exacerbate existing disparities [13], while privacy concerns, especially regarding data protection regulations like the GDPR, demand robust compliance.

These challenges and opportunities converge in the Vietnamese legal landscape. The feasibility of applying RAG to address the specific disparities in Vietnam remains an open question. RAG systems offer the potential for a cost-effective, scalable solution, yet they must be specifically tailored to the Vietnamese context. The ability of models like GPT to handle Vietnamese syntax, tone markers, and idiomatic expressions enhances their suitability. However, their success is contingent upon the integration of local legal datasets, such as those from a comprehensive national legal corpus. This paper seeks to bridge this research gap by evaluating the specific feasibility of such a system.

This paper proposes and evaluates a novel technological intervention: a legal assistance tool integrating RAG with the GPT LLM, tailored for Vietnam. The proposed research methodology comprises several key stages. First, a well-structured Vietnamese legal corpus, including statutes, case law, and educational materials, will be curated to enable effective retrieval. Second, the RAG model will be developed, pairing GPT's generative capabilities with a retrieval mechanism. Third, legal experts will validate the system's accuracy to ensure alignment with Vietnamese law. Finally, a continuous improvement loop will integrate user feedback and legal changes, ensuring the system remains a reliable resource.

2. MATERIALS AND METHODS

This study employed a developmental research design, centered on the construction and implementation of a Retrieval-Augmented Generation (RAG) system tailored for the Vietnamese legal context.

2.1 Model Selection

GPT was selected as the generative component due to its advanced conversational abilities and robust performance in zero-shot learning scenarios. This model's suitability is further enhanced by its strong capabilities in processing the Vietnamese language, including its complex syntax, tonal markers, and cultural idioms. To ensure a balance between natural-sounding responses and factual accuracy, the model was configured with a temperature of 0.7, max tokens limited to 1,000, and both frequency and presence penalties set to 0. This configuration allows the model to repeat key legal terms as necessary for clarity.

Furthermore, the system's architecture is based on the LangChain RAG pipeline, with specific customizations to address the unique challenges of Vietnamese legal text. The process begins with Query Preprocessing. To enhance clarity, user queries are automatically rewritten using a prompt template. This step resolves issues such as typos, abbreviations, and vague phrasing by instructing the model to act as a Vietnamese legal expert and rephrase the query for conciseness and precision.

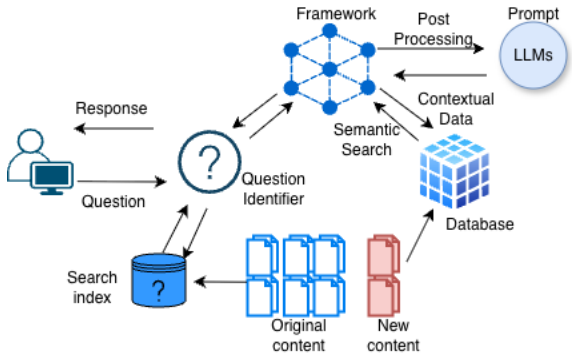

Figure 1. The proposal hybrid RAG framework

The core of the architecture is a hybrid document retrieval mechanism that combines traditional keyword-based search with advanced semantic search, as presented in Figure 1.

Step 1: Query Input The process begins when the Client submits a Question to the system's central hub, the Question Identifier.

Step 2: Query Triage (Hybrid Search) The Question Identifier analyzes the query to determine the best path forward. Path A (Keyword Search): For simple, direct queries, it can perform a traditional keyword search using the Search Index. The diagram implies it can return a simple "Response" from here. Path B (Semantic Search): For complex, nuanced questions, it passes the query to the Framework.

Step 3: Retrieval (Semantic Search) Once the Framework receives the query, it initiates a Semantic Search of the Vector Database to find documents that are conceptually and semantically similar to the question. This database is built from both "Original Content" and "New Content."

Step 4: Augmentation and Generation the Vector Database returns the relevant search results as Contextual Data back to the Framework. The Framework then "augments" the original query with this new context, creating a complete, data-rich Prompt. This prompt is then sent to the LLM.

Step 5: Post-Processing and Final Response The LLM generates its answer and sends it back to the Framework for Post Processing. This step might involve formatting, fact-checking, or simplifying the language. This final, processed response is sent to the Question Identifier, which delivers it to the Client.

Here, keyword search: The BM25 algorithm is employed for term-frequency-based retrieval, targeting exact keyword matches. And *vector search:* A Vietnamese-specific embedding model generates dense embeddings, which are stored in a vector database to facilitate semantic similarity search.

Scores from both BM25 and vector search are normalized and combined in a weighted manner to balance keyword precision with semantic relevance. This hybrid approach is enhanced by an Active Retrieval strategy; if the initial high-confidence results do not sufficiently address the query, the system iteratively explores lower-ranked documents to maximize recall.

Finally, the LangChain framework integrates these components. A retriever module creates a unified interface for the hybrid search outputs, passing the retrieved document chunks and their metadata to the GPT model. The LLM then synthesizes this grounded context to generate the final, accurate answer.

2.2 Dataset

Two distinct datasets were developed: a comprehensive retrieval corpus of legal articles and a benchmark dataset of question-answer pairs for evaluation.

2.2.1 Retrieval Corpus

A comprehensive legal corpus representing the Vietnamese legal system was curated, consisting of 500.000 Vietnamese legal articles from 23.000 unique legal documents. To optimize this corpus for retrieval, articles

were segmented into overlapping chunks (≤ 256 tokens), enriched with metadata, and processed for both BM25 (stopwords removed) and vector search (dense embeddings). This hybrid preparation allows the system to rank results based on both keyword and semantic similarity.

2.2.2 Evaluation Benchmark:

A benchmark dataset was created by scraping question-answer pairs from prominent Vietnamese legal websites (thuvienphapluat.vn, hieuluat.vn) using a custom Python tool. This dataset, covering diverse legal categories, was cleaned and mapped to single-document answers. During evaluation, these questions simulate user queries, allowing the system's generated answers and retrieved chunks to be compared against the ground-truth answers for accuracy.

2.3. Evaluation Methodology

To assess the system's performance, a qualitative evaluation was designed using a tiered, case-study approach. This framework was developed to test the system's retrieval accuracy, legal reasoning, and practical utility across a spectrum of legal complexity.

Three distinct categories of hypothetical cases, corresponding to simple, medium, and complex scenarios, were created in Table 1.

Table 1. Hypothetical case

Case	Scenario	Query	Summarized
Property Ownership Dispute (Simple/ Medium Complexity)	A dispute over a land sale agreement where one party (Mr. Nhan) alleges the contract was altered by the other (Mr. Long) to include additional land.	How should the Court determine the authenticity of the purchase agreement and the rightful ownership of the land?	The system correctly identified that the Court must evaluate the contract's authenticity under the Land Law 2013 and Civil Code 2015 by verifying notarized copies, checking cadastral records, and using forensic analysis.
Sale of Controlled Substances (Medium Complexity)	A pharmacy company was caught distributing controlled medications without a license, claiming a supply chain delay as the reason.	What penalties apply to unauthorized distribution of controlled substances, and can supply chain issues be considered a mitigating factor?	The system correctly retrieved penalties from the Penal Code 2015, (fines, imprisonment, corporate liability). It correctly identified that while mitigating factors exist (Article 51), "supply chain delays" are not explicitly recognized as one and do not absolve the company of its legal responsibility
Criminal Case – Self-Defense (Complex)	During an altercation, Mr. Khoa stabbed Mr. Tuan, resulting in Mr. Tuan's death. Mr. Khoa claims self-defense, but witness statements conflict.	What criteria should be used to determine whether Mr. Tuan acted in self-defense or committed murder, and how should witness testimony influence the ruling?	The system correctly identified the key legal articles (Article 22 for self-defense, Article 123 for murder) and the relevant legal tests: the presence of an imminent threat, the proportionality of the response, and the user's intent (premeditated or reactive).

3. RESULTS AND DISCUSSION

The quantitative results from the expert panel evaluation are summarized in Table 2. The overall average score across all seven metrics was 5.7, a strong positive result falling between "Somewhat Agree" (5) and "Agree" (6) on the 7-point scale.

Table 2. Expert Evaluation Scores by Case

Evaluation Metric	C1	C2	C3	Average
1. Response Accuracy	5.9	5.7	5.6	5.7
2. Clarity and Precision	5.5	5.1	5.6	5.4
3. Contextual Reasoning	6.1	5.7	6.5	6.1
4. Quality of Approach & Reasoning	6.6	6.2	6.5	6.4
5. Effectiveness in Addressing Query	5.3	6.8	4.7	5.6
6. Accuracy of Source Processing	5.8	5.5	5.9	5.7
7. Quality of Output Style	5.2	6.3	5.3	5.6
Per case average	5.5	5.8	5.9	5.7

The expert evaluations demonstrate strong approval of the system's core functionalities. The system's "approach and reasoning" received the highest commendation, with an average score of 6.7, indicating strong agreement from the

experts. This suggests the system's methodological foundation and logical structuring are its most successful aspects.

Scores for contextual reasoning, accuracy, and processing of legal acts were also high, confirming the RAG system's capability to retrieve and apply relevant legal information.

The "Per case average" remained highly consistent across simple, medium, and complex cases. This finding is significant, as it suggests the system's performance does not degrade when faced with increasing legal complexity. Notably, the system achieved its highest score for "approach and reasoning" on the most complex case (C3).

The lowest-rated items were "effectiveness" and "clarity and precision". While still positive, these scores are visibly lower than the others and suggest that while the system's reasoning is strong, its final output could be more concise and directly address the user's query.

The findings of this study affirm the potential of a RAG system, integrated with GPT-4, as a transformative tool for enhancing access to legal assistance in Vietnam. The positive expert evaluations, which resulted in an overall average score, indicate that the model performs effectively in retrieving relevant legal provisions and generating coherent, contextually accurate responses across a spectrum of legal complexities.

Interpretation of Key Findings

The study's results offer specific insights into the system's strengths and weaknesses. The highest-rated metric by a significant margin was the system's approach and reasoning, which achieved an average score (Table 2). This strongly suggests that the core technical choice a hybrid search mechanism combining semantic and keyword-based retrieval is highly effective. This hybrid approach allows the system to handle the linguistic subtleties of the Vietnamese legal context, where precision is critical.

Conversely, the system's lowest-rated (though still positive) metrics were its ability to address the question effectively and clarity and precision. This quantitative finding aligns with qualitative feedback from the legal experts. While the system's reasoning was sound, its final outputs were not always as concise or direct as they could be. This highlights a clear avenue for future refinement: focusing on prompt engineering that instructs the model not only to be accurate but also to be succinct.

Limitations and Operational Considerations

While the system's performance was consistent across simple, medium, and complex cases (Table 2), expert feedback identified a key limitation. The system's ability to handle highly complex scenarios, such as jurisdictional conflicts or reconciling overlapping and contradictory legal frameworks, requires further refinement. This reflects a known challenge in computational law, as such tasks demand advanced reasoning and domain-specific training that pushes the boundaries of current-generation LLMs.

From an operational perspective, this study underscores that the system's reliability is contingent upon its knowledge base. The RAG system's accuracy is a direct function of its non-parametric (retrieved) data. Therefore, maintaining a well-curated legal corpus, validated by experts and continuously updated, is essential for its long-term relevance. Furthermore, while the system demonstrated proficiency in Vietnamese, expert feedback suggested that additional training on regional dialects and specialized legal terminologies would further enhance its accessibility and inclusivity.

Ethical Implications and Governance

The deployment of an AI-powered legal assistance tool necessitates a robust discussion of its ethical implications. Key issues include data privacy, the potential for algorithmic bias inherited from training data, and the critical need to manage user expectations. The system must be unambiguously positioned as a legal guidance tool, not a substitute for definitive legal advice from a qualified professional.

Ensuring the system aligns with the principles of justice and fairness requires collaboration between policymakers, legal professionals, and community organizations. Clear guidelines and regulations must be established to govern its use, particularly in a high-stakes domain like law.

Future Research Directions

Based on these findings, future research on RAG for legal systems should focus on three key areas:

Interpretability: Improving the trust and transparency of RAG outputs is critical. Implementing explainable AI (XAI) techniques could help users understand the rationale behind the system's recommendations, which is essential in high-stakes legal scenarios [14].

Accessibility: Enhancing multilingual capabilities and, more specifically, incorporating a wider range of regional dialects will be crucial for expanding the system's accessibility to all diverse populations within Vietnam.

Governance: Further research is needed to develop robust ethical and governance frameworks for AI in law. This will be essential to ensure that these technologies contribute to a more equitable and inclusive legal system.

4. CONCLUSION

This study validates the feasibility of a Retrieval-Augmented Generation (RAG) system, powered by GPT-4, to effectively address Vietnam's legal literacy gap. Strong performance metrics and positive expert evaluations confirm the system's practical utility and its potential to democratize access to justice for marginalized populations. While this research provides a proven framework, future work must focus on enhancing the system's capacity for complex legal reasoning, expanding its training corpora, and establishing robust ethical governance. This study contributes actionable insights for developing AI-powered legal assistance in emerging economies, offering a replicable model for more inclusive and equitable public service.

5. REFERENCES

[1] Nguyen Quoc Suu. Law dissemination and education for ethnic minorities in mountainous areas of

Vietnam, Nuances , vol. 33, no. 00, p. e022024, Dec. 2022.

DOI: 10.32930/nuances.v33i00.9724.

[2] Hang, P. T., et al. Assessment of the impact of ethnic minorities' capacity in Binh Dinh province, Vietnam, to access land and exercise land use rights by application of linear structure model (PLS-SEM). In IOP Conference Series: Earth and Environmental Science, 1403(1), p. 012005. IOP Publishing, 2024.

[3] Ehlert, Judith, and Martina Padmanabhan. The missing lens: feminist perspectives on agrofood system transformation in Vietnam and Indonesia. Österreichische Zeitschrift für Soziologie, 35, pp. 50.1, 2025

[4] Ngan, Tran Thanh. Models and Methods of Dissemination and Legal Education of Legal Aid Centers in Ho Chi Minh City. Science and Education, 3(9), pp. 1946-1950, 2019.

[5] Pham, Hong Thai. The Issues of Justice and the Right of Access to Justice in Vietnamese Law. VNU Journal of Science: Legal Studies, 36(1), 2020.

[6] Gao, Yunfan, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2(1), 2023.

[7] Sahoo, Pranab, et al. A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927, 2024.

[8] Fan, Wenqi, et al. A survey on rag meeting LLMs: Towards retrieval-augmented large language models. Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. 2024.

[9] Surden, Harry. Artificial intelligence and law: An overview. Ga. St. UL Rev. 35, pp. 1305, 2018.

[10] Lai, Jinqi, et al. Large language models in law: A survey. AI Open 5, pp.181-196, 2024.

[11] Zhelyazkova, Asya. Challenges in EU law enforcement and the digital age. Research Handbook on the Enforcement of EU Law, pp. 91-105, 2023.

[12] Ariai, Farid, et al. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. arXiv preprint arXiv:2410.21306, 2024.

[13] Prem, Erich. From ethical AI frameworks to tools: a review of approaches. AI and Ethics, 3(3), pp. 699-716, 2023.

[14] Dwivedi, Rudresh, et al. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM computing surveys, 55(9), pp.1-33, 2023.