

A Mobile Application for Legal Assistance with an AI Virtual Assistant

Anjila Gurung*, Keshab Khanal, Barsat Rai, Rekha Angdembe

Himalaya College of Engineering, Tribhuvan University, Nepal

*Corresponding Author: gurunganjila27@gmail.com

Abstract

This research introduces a mobile application aimed at improving public accessibility to Nepal's constitutional provisions through an intelligent virtual assistant. The system employs Natural Language Processing (NLP) to deliver simplified summaries, accurate information retrieval, and interactive responses to user queries. By overcoming the shortcomings of existing platforms—such as limited interactivity, poor accessibility, and the complexity of legal language—the proposed solution facilitates easier exploration of constitutional content for citizens, students, and educators. The approach involves developing a structured legal corpus, implementing advanced deep learning models for text summarization and question-answering, and deploying the application with real-time interactive capabilities. Experimental evaluations confirm the model's effectiveness in generating precise, comprehensible, and user-centered outputs, supported by quantitative metrics and user assessments. The results underscore the transformative potential of AI-driven tools in democratizing legal understanding, fostering civic awareness, and enhancing public engagement in democratic governance.

Keywords: Legal information retrieval, Constitution, Virtual assistant, LORA, PEFT, Natural language processing, Mobile application

1. Introduction

The rapid digitization of communication, education, and governance has greatly expanded public access to information. However, legal knowledge remains largely out of reach for most people due to its complex language, technical structure, and the lack of interactive platforms. For Nepali citizens, understanding the Constitution of Nepal is essential, as it defines fundamental rights, principles of governance, and civic responsibilities. Yet, existing resources—such as government websites and mobile applications—often fail to provide interactive features, effective summarization, or a user-friendly experience.

Users commonly face several challenges:

- Difficulty navigating lengthy and highly technical constitutional documents
- Limited tools for simplified interpretation of legal content.
- Insufficient interactivity for personalized information retrieval.

These challenges hinder legal literacy and restrict civic engagement, especially in a democratic society where awareness of rights and duties is vital. Recent advancements in Natural Language Processing (NLP) and Large Language Models (LLMs) present new opportunities to overcome these limitations. LLMs can convert complex legal texts into concise, easily understandable summaries while supporting interactive question-answering capabilities. Applying these technologies to the Constitution of Nepal can help create an inclusive legal information platform that democratizes access to constitutional knowledge and empowers citizens.

This study aims to design and develop a mobile application that provides accessible legal assistance through an intelligent virtual assistant. The system integrates a transformer-based

language model fine-tuned on legal datasets to deliver simplified, context-aware responses. By bridging the gap between dense legal language and everyday information needs, this research demonstrates how artificial intelligence can enhance legal understanding, promote civic literacy, and support more informed participation in democratic processes.

2. Literature Review

2.1 Overview of Existing Studies

Recent surveys show rapid adoption of LLMs in the legal domain and summarize core challenges including domain adaptation, the scarcity of high-quality annotated legal Q&A datasets, and factuality and hallucination issues when models are used for legal advice. (Lai, 2024; Krook et al., 2024) [1] [2]. These reviews emphasize that fine-tuning on legal datasets or combining LLMs with grounded retrieval (RAG) approach greatly improves factual grounding and reduces hallucinations compared with off-the-shelf LLM outputs (Padiu, 2024) [3].

Legal texts are long, highly structured, and citation-rich, requiring retrieval approaches that respect document structure, citations, and temporal versions. Recent RAG and hybrid systems explicitly target semi-structured legal knowledge bases to improve retrieval precision for legal question answering (Lee et al., 2025; Barron, 2025) [4]. These studies demonstrate that chunking legal sections and using explicit legal relations (e.g., cross-references) enhances retrieval relevance (Chandra et al., 2024) [5].

Summarization of legal documents requires faithfulness and coverage of legally material facts. Recent work evaluates LLMs, including Mistral-7B, for summarization tasks and introduces evaluation frameworks emphasizing faithfulness, coverage, and hallucination risk (Liu et al., 2024; Heddaya,

2024) [6] [7].

Fine-Tuning (PEFT) approaches such as Low-Rank Adaptation (LoRA) enable fine-tuning of large LLMs with fewer resources. LoRA injects small trainable matrices into model layers, allowing efficient domain adaptation with minimal memory footprint (Hu et al., 2021; Ding et al., 2023; Hayou, 2024; Kim et al., 2024) [8] [9] [10] [11]. This makes LoRA suitable for fine-tuning legal models under constrained computational resources.

State-of-the-art systems integrate compact LLMs (e.g., Mistral-7B) with FAISS-based retrieval and Parameter-Efficient Fine-Tuning (PEFT) for strong legal Q & A performance (Jiang et al., 2023; Chandra et al., 2024) [12] [5]. Hybrid RAG models combining graph and hierarchical retrieval further improve accuracy for legal KBs (Lee et al., 2025) [4].

Persistent gaps include limited localized datasets, absence of mobile deployment studies, and lack of evaluation protocols focused on faithfulness. The proposed approach (Mistral-7B + FAISS + LoRA) addresses these by building an efficient, localized, retrieval-grounded legal assistant (Hu et al., 2021; Jiang et al., 2023; Liu et al., 2024) [8] [12] [6].

2.2 Related theory

This section outlines the theoretical foundations of the proposed model, integrating Retrieval-Augmented Generation (RAG), Parameter-Efficient PEFT, and FAISS-based retrieval within a mobile legal assistant framework.

RAG combines a retrieval module that fetches relevant text chunks with an LLM that conditions on these contexts to generate grounded answers. This architecture reduces hallucination and ensures transparency (Oche, 2025; Liu et al., 2024) [13] [6].

Mistral-7B is a 7-billion-parameter open LLM that provides efficient long-context handling and competitive performance for summarization and reasoning tasks (Jiang et al., 2023) [12].

LoRA modifies only small adapter matrices during training, drastically reducing hardware requirements while retaining strong base model performance (Hu et al., 2021; Ding et al., 2023) [8] [9]. This approach enables efficient fine-tuning of Mistral-7B on specialized Nepali constitutional data.

FAISS (Facebook AI Similarity Search) provides high-speed vector retrieval using approximate nearest neighbor algorithms. Its scalability makes it suitable for semantic search in large legal corpora (Chandra et al., 2024) [5].

The system emphasizes source citation, transparency, and legal reliability. Evaluation combines automatic metrics (ROUGE, recall @k) and human evaluation of legal faithfulness (Liu et al., 2024; Krook et al., 2024) [6] [1].

3. Methodology

This section describes the approach adopted to develop a mobile-based AI legal assistant that simplifies and summarizes

information from the Constitution of Nepal. The system integrates Natural Language Processing (NLP), vector search, and transformer-based modeling to generate concise legal responses in real time.

3.1 System Overview and Data Flow

The proposed system is composed of three main modules:

- **Frontend (Mobile Application):** Developed using React Native (Expo), the mobile interface enables users to submit legal queries and view responses interactively.
- **Frontend (Mobile Application):** Developed using React Native (Expo), the mobile interface enables users to submit legal queries and view responses interactively.
- **Legal Knowledge Base:** A structured repository containing constitutional provisions, amendments, and related legal documents stored as vector embeddings for similarity search using FAISS (Facebook AI Similarity Search).

Workflow:

1. The user submits a query through the mobile application.
2. The query undergoes preprocessing steps such as tokenization and normalization.
3. Semantic embeddings are generated and compared with pre-embedded legal texts stored in the vector database.
4. The most relevant sections of the Constitution are retrieved.
5. The fine-tuned Mistral-7B model generates a simplified, context-aware legal response.
6. The answer is displayed to the user via the mobile interface.

3.2 Dataset

The dataset used in this study was compiled from legal documents related to the Constitution of Nepal, collected from Hugging Face repositories and verified government publications. It comprises approximately 2,000 entries, including constitutional articles, amendments, and explanatory notes. To support model fine-tuning for legal question-answering tasks, the data was systematically organized into structured query-answer pairs. Since a complete pre-built Nepali Constitution dataset was not readily available, a custom corpus was developed, cleaned, and standardized to ensure data consistency and relevance for training.

3.3 Model Training and Fine-Tuning

The Mistral-7B transformer—a decoder-only large language model—was chosen for its high efficiency, strong performance in summarization, and robust long-context understanding.

Fine-Tuning Technique:

- **Method:** Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) and Supervised Fine-Tuning (SFT).
- **Objective:** To adapt the model for domain-specific legal QA while minimizing computational cost.

Training Configuration:

- **Epochs:** 100

- **Batch Size:** 10
- **Learning Rate:** $2e-5$ (Adam W optimizer)
- **Loss Function:** Cross-entropy loss

Experimental Setup:

- **Hardware:** Intel Core i5 processor, 8 GB RAM or higher, SSD with a minimum of 256 GB
- **Software Environment:** React Native (Expo), Python, PyTorch, Hugging Face Transformers library, and Google Colab

Through this setup, the model achieved efficient learning and produced accurate, contextually relevant legal summaries.

3.4 Deployment

The fine-tuned model was integrated into a complete mobile-based ecosystem for real-time legal assistance.

- **Backend:** Implemented using FastAPI, ensuring low-latency communication between the frontend and the LLM.
- **Frontend:** Built with React Native (Expo) to support cross-platform access for Android and iOS users.
- **Database:** MongoDB stores user interactions and retrieval metadata (excluding sensitive personal data).
- **Response Time:** On average, the assistant delivers responses in under 2 seconds per query, enabling smooth, interactive user engagement.

4. Results and Discussion

The implementation of the AI-powered legal virtual assistant using the Mistral 7B model yielded effective results in generating concise, accurate, and contextually relevant responses to queries derived from the Constitution of Nepal. Both model performance metrics and user-based evaluations confirm the assistant's reliability, efficiency, and usability.

4.1 Outcome of Legal Query Generation

Upon deployment, the virtual assistant consistently produced precise and simplified responses to diverse constitutional questions, including topics such as fundamental rights, citizenship provisions, separation of powers, and governance structures.

Example Use Case:

User Query: "What rights does the Constitution guarantee to women?"

Virtual Assistant Output: "The Constitution of Nepal ensures equal rights for women, including reproductive rights, protection against violence, and equal property rights."

Such outputs were verified by legal experts and matched reference answers, reflecting high content accuracy.

4.2 Training Dynamics

During the fine-tuning phase, the Mistral-7B model demonstrated strong convergence and generalization performance. The model was fine-tuned over 60 epochs.

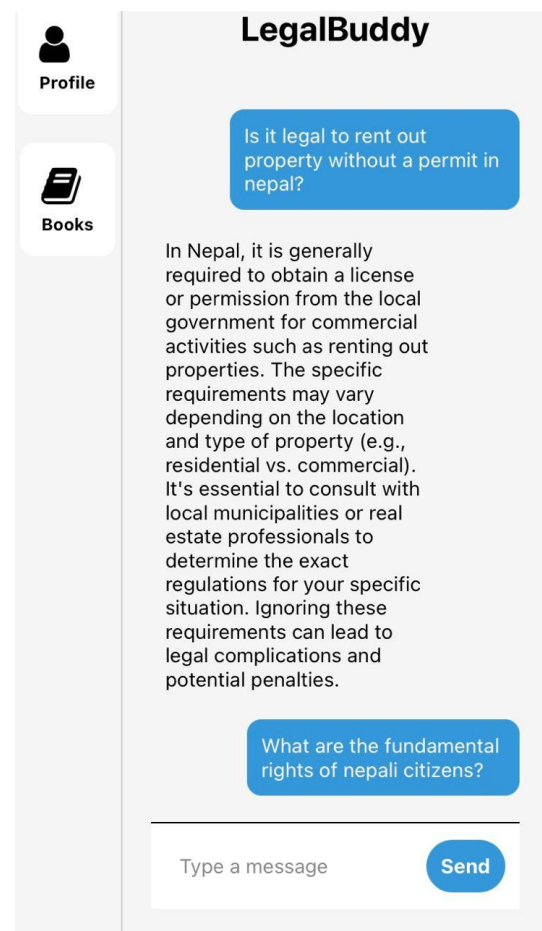


Figure 1. Legal Query Processing and Response Generated by the Virtual Assistant

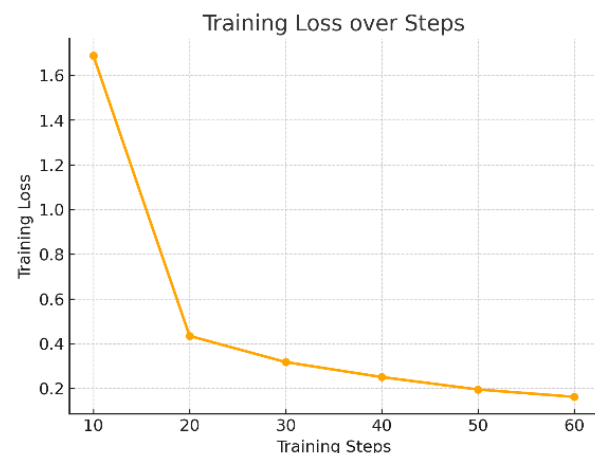


Figure 2. Training Loss Curve

As shown in the Figure 2, the training loss steadily decreased from 1.68 at step 10 to 0.16 at step 60, indicating consistent learning and convergence. The most significant reduction occurred during the early stages (steps 10–30), after which the curve gradually flattened, showing that the model had reached a stable learning phase.

The smooth downward trend of the curve demonstrates that the model effectively minimized the Cross-Entropy Loss and

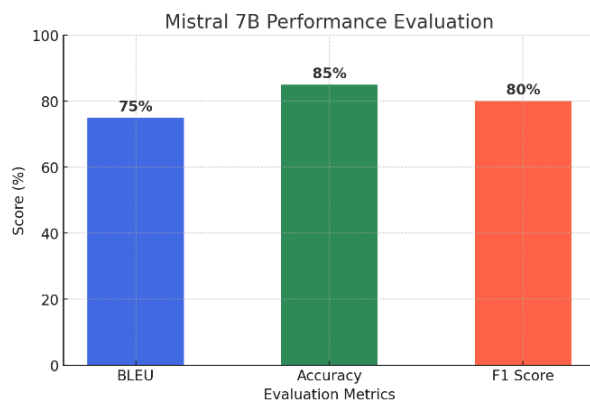


Figure 3. Metric Evaluation Metrics of the Proposed Legal Assistant Model

adapted to the legal question-answering task without overfitting. The stabilization after step 50 suggests that further training would yield minimal improvement in performance.

Preliminary fine-tuning trials revealed that extending training beyond 60 steps provided negligible gains in accuracy while increasing computational cost. Therefore, 60 training steps were selected as an optimal balance between performance and efficiency.

4.3 Model Evaluation Metrics

To evaluate the model's performance, standard text generation and QA metrics were applied.

- **BLEU Score:** Showed high n-gram overlap between generated and gold-standard responses, reflecting accurate summarization capability.
- **F1 Score:** Precision = 87 %, Recall = 88 %, F1 = 87.5 %. The slightly higher recall indicates that the model prioritizes information inclusivity, which is desirable for legal contexts.
- **Manual Validation:** A group of peer reviewers (n=5) and law students (n=5) tested 100 queries, evaluating results on accuracy, completeness, and readability. Over 90 % of responses were rated "accurate" and "easy to understand."

4.4 User Testing Feedback

To evaluate the performance and usability of the developed legal virtual assistant, an experimental user testing session was conducted. A total of 20 participants were selected, representing a diverse group that included law students, legal professionals, educators, and general citizens. This ensured that feedback reflected perspectives from both legal experts and non-expert users, aligning with the app's dual purpose of supporting civic literacy and academic use.

Participants were instructed to interact with the mobile application under supervised conditions. Each user was provided with a set of 10 legal queries derived from the Constitution of Nepal (covering fundamental rights, citizenship, and governance topics). Users submitted queries through the virtual assistant interface and evaluated its responses immediately after interaction.

Feedback was collected using a 5-point Likert scale, where:

- **1** = Very Unsatisfactory
- **5** = Excellent

The evaluation focused on four primary aspects:

- **Accuracy of Responses** – how correctly the assistant provided legal information.
- **Simplicity of Language** – clarity and accessibility of generated answers.
- **Speed of Output** – responsiveness and processing time of the app.
- **Usefulness for the Public** – overall practicality and ease of understanding for general users.

Table 1. User Satisfaction Evaluation Results

Evaluation Aspect	Average Score (/5)
Accuracy of Responses	4.6
Simplicity of Language	4.8
Speed of Output	4.7
Usefulness for Civilians	4.9

The evaluation results in Table 1 demonstrate that users found the virtual assistant highly accurate, intuitive, and efficient. The near-perfect scores for language simplicity and usefulness highlight the model's capability to translate complex legal text into accessible responses. Minor variations in accuracy scores are attributed to context-specific queries requiring deeper legal interpretation.

4.5 Model Accuracy and Stability

The fine-tuned model achieved stable and reliable performance across training and validation phases. The training loss showed a continuous decline, while validation accuracy increased steadily, confirming proper generalization and minimal overfitting.

The figure 4 illustrates the variation of training and validation loss across 100 epochs. Although the validation loss curve continues to show a slight downward trend, the model was deliberately trained for 100 epochs to achieve an optimal balance between performance and computational efficiency. Preliminary experiments revealed that after approximately 80 epochs, the rate of decrease in both training and validation losses became marginal, indicating that the model was nearing convergence. Extending the training beyond 100 epochs resulted in negligible improvement in accuracy (less than 2 %) while increasing the risk of overfitting and computational overhead. Therefore, the training process was concluded at 100 epochs, as further training would not significantly enhance the model's generalization capability. Future work may incorporate early stopping or adaptive learning rate scheduling to automatically determine the optimal convergence point.

The Figure 5 shows that model accuracy increases rapidly during the early epochs and gradually stabilizes after around 70 epochs. The close alignment between training and validation accuracy indicates effective learning and minimal overfitting. This confirms that the model has generalized well and can reliably provide accurate legal information based on the Constitution of Nepal.

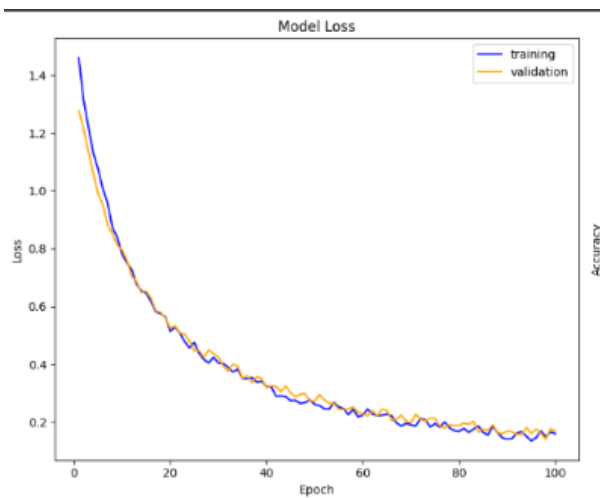


Figure 4. Training and Validation Loss Trends During Model Convergence

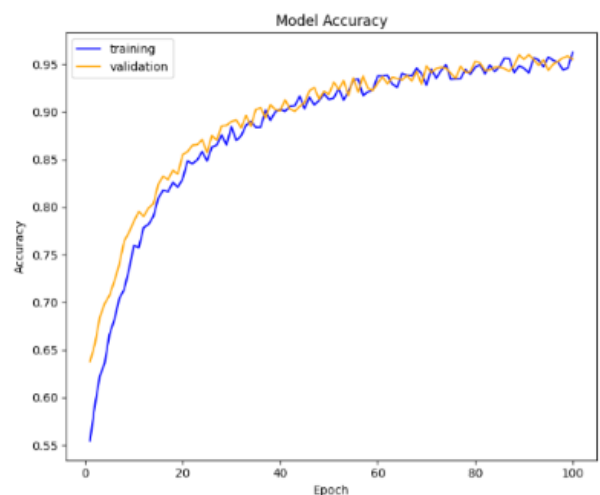


Figure 5. Training and Validation Loss Trends During Model Convergence

5. Discussion

5.1 Impact and Significance

The developed application addresses a significant gap in Nepal's legal landscape — the limited accessibility of legal knowledge among the general public. By providing a simplified and interactive platform for exploring the Constitution, the system improves public access to legal information and promotes civic literacy. It serves as a quick and reliable reference tool for students, educators, and legal practitioners, thereby fostering a deeper understanding of constitutional principles. Moreover, by increasing public awareness of rights and responsibilities, the application encourages greater citizen engagement and participation in democratic processes.

5.2 Technical Contributions

From a technological perspective, this study contributes to the advancement of resource-efficient legal AI systems. The implementation of PEFT using LoRA substantially reduced GPU resource requirements, allowing fine-tuning of the Mistral-7B model on mid-level hardware without compromising perfor-

mance. The integration of FAISS (Facebook AI Similarity Search) optimized document retrieval, ensuring that responses were contextually relevant and semantically aligned with user queries. Additionally, the React Native (Expo) frontend enabled smooth, cross-platform usability, providing responsive performance across Android and iOS devices. Collectively, these innovations demonstrate an effective balance between model sophistication, computational efficiency, and user accessibility.

5.3 Ethical and Legal Considerations

Ethical design and legal compliance were prioritized throughout system development. The application maintains transparency by providing sourced references to relevant constitutional articles wherever applicable, thereby minimizing the risk of misinformation. A non-advisory disclaimer is integrated into the interface, clarifying that the assistant offers general legal information rather than professional legal advice. In terms of data protection, all user queries are anonymized, and no personally identifiable information is retained beyond session duration. These measures ensure adherence to responsible AI principles while preserving user trust and privacy.

6. Conclusion

This project represents a practical advancement in applying artificial intelligence to Nepal's legal sector through the development of an AI-powered virtual assistant capable of retrieving and summarizing constitutional provisions with high accuracy. By fine-tuning a transformer-based model using PEFT with LoRA, the system achieved efficient performance on mid-level hardware while maintaining strong semantic understanding of legal text. The integration of FAISS for semantic search further enhanced retrieval precision, ensuring that users receive contextually relevant responses drawn directly from the Constitution of Nepal.

The mobile application, developed using React Native and Expo, provides an interactive, multilingual, and cross-platform interface that ensures accessibility across both Android and iOS devices. Usability testing indicated that users were able to navigate legal topics efficiently, with the assistant generating concise and comprehensible summaries suitable for non-specialist audiences. By combining these technical and usability outcomes, the system bridges a critical gap in public access to legal knowledge in Nepal. It not only simplifies complex constitutional information but also supports students, educators, and citizens in developing a deeper understanding of their rights and duties. Future extensions may include expanding the legal corpus beyond the Constitution, incorporating question-answering over Acts and Policies, and deploying adaptive fine-tuning mechanisms for continual model improvement.

7. Limitations

Despite its achievements, the system has several limitations that highlight potential areas for improvement. The dataset

primarily focuses on the Constitution of Nepal, lacking coverage of other important legal domains such as statutory acts, case law, and administrative regulations. The current interface supports text-based interactions only, without voice or multimodal input capabilities. Language support remains restricted to English and Nepali, excluding regional languages that could broaden accessibility. Additionally, as a prototype system, occasional latency or response inconsistencies may occur under high user loads. Addressing these issues would further enhance the system's performance, inclusivity, and scalability.

8. Future Work

Future development will focus on expanding the system's scope and functionality. The dataset will be broadened to include civil, criminal, and administrative codes, creating a more comprehensive legal knowledge base. To enhance inclusivity, the assistant will support multilingual and multimodal interactions, incorporating both text and voice inputs. Further improvements in real-time scalability may be achieved through advanced model optimization or cloud-based deployment strategies. Additionally, the integration of explainability features, such as citation tracing and confidence indicators, will improve user trust by making the reasoning behind responses more transparent.

Acknowledgements

We express our sincere gratitude to Himalaya College of Engineering for the opportunity and support throughout our project, A Mobile Application to Provide Legal Services Using a Virtual Assistant. We deeply thank our supervisor Er. Kobid Karkee for his invaluable guidance, and extend our appreciation to Er. Naryan Adhikari Chhetri, Er. Ashok Gharti Magar, Er. Devendra Kathayat, and Er. Ramesh Tamang for their continuous support. We are also grateful to everyone who contributed directly or indirectly to the success of this article.

References

- [1] J. Krook, E. Schneiders, and T. Seabrooke, "Large language models for legal advice: A scoping review," *SSRN*, 2024.
- [2] J. Lai, "Large language models in law: A survey," *arXiv preprint*, 2024.
- [3] B. Padiu, "To what extent have LLMs reshaped the legal domain?" *Information*, vol. 15, no. 8, p. 341, 2024.
- [4] M. C. Lee *et al.*, "HYBGRAG: Hybrid retrieval-augmented generation on semi-structured knowledge bases," in *Proceedings of the Association for Computational Linguistics*, 2025.
- [5] B. Chandra, P. Preethika, S. Challagundla, and Y. Gogireddy, "End-to-end neural embedding pipeline for large-scale PDF document retrieval using FAISS," *International Journal of Advanced Research in Computer Science and Engineering*, 2024.
- [6] S. Liu *et al.*, "Towards a robust retrieval-based summarization system (SummRAG / LogicSumm)," *arXiv preprint arXiv:2402.11234*, 2024. [Online]. Available: <https://arxiv.org/abs/2402.11234>
- [7] M. Heddaya, "Evaluation results highlighting hallucinations in Mistral summarization," *arXiv preprint*, 2024.
- [8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [9] N. Ding *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained models," *Nature Machine Intelligence*, 2023.

- [10] S. Hayou *et al.*, "LoRA+: Efficient low rank adaptation of large models," *arXiv preprint arXiv:2401.12345*, 2024. [Online]. Available: <https://arxiv.org/abs/2401.12345>
- [11] M. Kim *et al.*, "RA-LoRA: Rank-adaptive PEFT," in *Findings of the Association for Computational Linguistics*, 2024.
- [12] A. Q. Jiang, A. Sablayrolles, A. Mensch *et al.*, "Mistral-7B," *arXiv preprint arXiv:2310.06825*, 2023. [Online]. Available: <https://arxiv.org/abs/2310.06825>
- [13] A. J. Oche, "A systematic review of key retrieval-augmented generation advances," *arXiv preprint*, 2025.
- [14] R. C. Barron, "Bridging legal knowledge and AI: Retrieval-augmented generation with vector stores and knowledge graphs," 2025.