NepKanun: A RAG-Based Nepali Legal Assistant

Abstract

Accessing legal information in Nepal is difficult due to complex terminology, limited resources, and misinformation . We introduce an AI-powered legal assistant that is tailored for Nepali legal texts and is built on a fine-tuned large language model. The technology provides precise, streamlined answers to natural language legal inquiries when integrated into a Retrieval-Augmented Generation (RAG) framework. It was trained using a custom dataset of high-quality question-answer pairs, and according to BERTScore, it obtained strong F1 scores of 0.82 (simple), 0.77 (moderate), and 0.71 (complex). Its usability is further confirmed by expert reviews. Our method shows how merging generation and retrieval can effectively democratize access to legal knowledge in Nepal by focusing on customized legal data and incorporating RAG.

Keywords: Large Language Model, Retrieval-Augmented Generation, Parameter-Efficient Fine-Tuning, Natural Language Processing, Low-Rank Adaptation, Optical Character Recognition

1 Introduction

The intersection of Artificial Intelligence (AI) and law presents transformative opportunities to enhance access to justice and legal knowledge globally (Alarie et al., 2016). However, people speaking low-resource languages like Nepali have suffered greatly as a result of these developments, which have disproportionately benefited high-resource language regions. The complexity of legal terminology and structures, along with low public knowledge and limited access to judicial resources, frequently makes it difficult for people in Nepal to grasp their legal rights and duties.Although NLP has advanced information retrieval and question answering in various domains, its application to Nepali legal texts is constrained by the scarcity of annotated datasets and the specialized nature of legal language (Paudel et al., 2024). The urgent need for customized AI solutions is highlighted by the fact that Nepal's current legal information systems frequently rely on keyword matching and lack the semantic depth necessary to correctly comprehend user inquiries.

Large Language Models (LLMs), such the Llama series (Touvron et al., 2023; Grattafiori et al., 2024), GPT-4 (OpenAI et al., 2024), and others (Zhao et al., 2023), have greatly improved natural language processing (NLP), allowing for more complicated task-solving capabilities and more natural human-computer interaction . Despite their ability in understanding and generating text, even in languages like Nepali (Duwal et al., 2024) , LLMs may display factual errors or "hallucinations," particularly in fields that need a lot of expertise or are changing quickly . To address these problems, Retrieval-Augmented Generation (RAG) has surfaced, which combines LLMs with an external retrieval step and grounds responses in reliable sources prior to generation. (Lewis et al., 2020). This procedure enables the integration of domain-specific knowledge and establishes the output in validated facts, greatly enhancing accuracy and traceability, particularly for knowledge-intensive tasks (Gao et al., 2024).Since RAG provides contextually appropriate responses, its application has demonstrated promise in enhancing the dependability of legal information systems (Ryu et al., 2023).

Addressing the challenges of accessing legal information in Nepal, this paper's key contributions are:

- Creation of a domain-specific, superior dataset that includes Nepali legal question-answer pairs in order to fill the gap in annotated legal data.
- Design and implementation of a RAG-based system that generates precise, context-aware

answers to legal queries by utilizing an LLaMA 3.2 3B model that has been refined on a unique Nepali legal QA dataset.

2 Related Work

The goal of improving the accessibility and manageability of legal information has led to significant advancements in the application of NLP techniques in the legal field. Legal text NLP tasks have been greatly influenced by transformer-based models, which include foundational models like BERT (Devlin et al., 2019) and subsequent large language model variants (OpenAI et al., 2024).

Efforts have been concentrated on adapting these general-purpose models to the particular character of legal language. For instance, the Legal-BERT family was created by Chalkidis et al. (2020) and pre-trained on a variety of legal corpora from the US, UK, and European countries. This domain-specific pre-training improved performance on downstream legal tasks such as classification and named entity recognition. However, its reliance on English data limits its applicability to non-English legal systems.Work in non-Western contexts, such as (Paul et al., 2023), pre-training on Indian court rulings , emphasizes the importance of adapting models to specific legal systems and languages.

In the context of Nepali, which is considered a low-resource language, foundational linguistic work by (Bal, 2004) has been essential for comprehending the structure of the language and facilitating the development of fundamental NLP tools.Efforts have focused on developing general-purpose Nepali language models, such as NepBERTa (Timilsina et al., 2022) trained on large monolingual corpora, and investigations into different Transformer-based models for Nepali text classification (Maskey et al., 2022) . These works represent significant steps in building foundational NLP capabilities for Nepali.However, direct applications to the complex Nepali legal domain remain limited. A notable contribution is by (Poudel et al., 2024), who developed a transformer-based bidirectional Neural Machine Translation system for English-Nepali legal texts, alongside a pioneering parallel corpus of 125,000 sentences. Their work demonstrates the feasibility of transformers in legal NLP for Nepali, while highlighting the challenge of data scarcity for question answering and simplification for the general public.

Despite these advancements, easily accessible and intelligible legal information for Nepali speakers remains scarce. Our work builds on prior contributions by creating an AI-powered legal question-answering system that combines RAG with a Nepali law-specific LLM.

3 Methodology

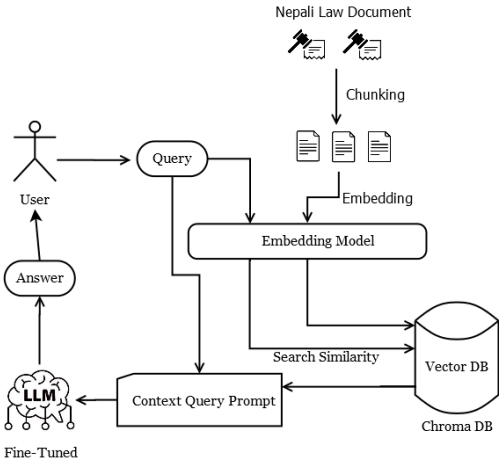

Figure 1: Overview of the workflow for AI assistance in Nepali law

3.1 Data Preparation

A high-quality, domain-specific dataset was necessary to modify the LLM to answer Nepali legal questions. We began curating a dataset with over 16,000 entries gathered using a hybrid strategy that involved online scraping from the Supreme Court website and pertinent news sources as well as processing PDF legal documents.To align with common user queries, these texts were transformed into question-answer pairs specifically designed for instruction tuning using prompt engineering. PyTesseract for Devanagari script (Paudel et al., 2024) was used to digitize the scanned documents, and they underwent essential manual validation. A systematic data cleaning pipeline, involving deduplication and correction of OCR-induced errors, yielded the final 10,000 high-quality entries utilized for model fine-tuning.

3.2 Fine-Tuning

We selected the Llama 3.2 3B instruct model (Meta AI, 2024), noting its balance of model size, performance characteristics, and suitability for multilingual instruction-following tasks. We employed Parameter-Efficient Fine-Tuning (PEFT), specifically Low-Rank Adaptation (LoRA) (LoRA) (Hu and et al., 2021) and Quantized LoRA (QLoRA)

(Dettmers and et al., 2023). Low-rank matrices are introduced by LoRA to decrease trainable parameters, and the base model is quantized to 4-bit precision by QLoRA to further increase efficiency. The Unsloth library (Han and Schick, 2023) was used to implement these strategies, which were tuned for quicker LoRA fine-tuning and lower memory consumption.

3.2.1 Training Configuration

The fine-tuning process utilized the following hyperparameters:

Parameter	Value
Learning Rate	3e-4 and 2e-4
Weight Decay	0.01
Scheduler	Cosine
Optimizer	AdamW (8-bit)
Mixed Precision	BFloat16
Max Sequence Length	2048 tokens

Table 1: Hyperparameters for fine-tuning LLaMA 3.2 3B

The LoRA rank was set to 16, updating approximately 24 million parameters. Using L40s GPUs, the training was carried out in the Lightning AI Studio (Lightning AI, 2024).

3.3 RAG Methodology

We used a Retrieval-Augmented Generation (RAG) architecture to make sure the produced responses are correctly based on reputable legal materials (Lewis et al., 2020). RAG systems are especially well-suited for fields like law, where having access to accurate, validated data from a particular corpus is crucial (Gao et al., 2024). The framework combines a generative component (the refined LLM) that synthesizes the final response based on the recovered context and the user query with a retrieval component that finds relevant data chunks from a knowledge base.

3.3.1 Knowledge Base Construction and Preprocessing

Important Nepali legal texts, such as the Constitution of Nepal 2072 and portions of important legislation (such as the Environmental Act and Muluki Ain), served as the main source of information for the RAG system. PyTesseract for OCR was used to handle a large number of documents that were available as scanned PDFs. For efficiency, a caching mechanism was included. We used a structure-aware chunking technique that divided text according to these logical divisions in order to maintain the semantic structure present in legal documents, which are frequently arranged hierarchically (e.g., into Parts, Chapters, and Articles).

3.3.2 Vector Embedding and Indexing

A multilingual SentenceTransformer model, based on architectures like Sentence-BERT (Reimers and Gurevych, 2019), was used to transform processed text chunks into dense vector embeddings. This model was selected because it can capture semantic linkages across languages, including Nepali. A vector embedding of 384 dimensions was used to represent each piece. ChromaDB (Contributors, 2023), an open-source vector database designed for effective similarity search, was used to index and store these embeddings.

3.3.3 Retrieval and Generation

Initially, a user's query is embedded into the same 384-dimensional vector space. The retrieval component does a similarity search to find the most relevant ChromaDB pieces. To enhance contextual coverage, we used Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to choose the top $k = 9$ document chunks, weighing variety across the retrieved sections against relevance to the query. The generating component, the fintuned Llama 3.2 3B model, receives the retrieved document chunks concatenated with the initial user query. The LLM incorporates data from the given context to provide the final response. The model's prompts and queries were carefully developed in Nepali to guarantee useful and contextually relevant results.

4 Results and Discussion

4.1 Automated Evaluation

Objective metrics were used to assess output quality. While ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) are common, their reliance on surface-level token overlap limits their effectiveness for evaluating semantic correctness, particularly in legal settings where paraphrasing can convey the same meaning. Therefore, we used BERTScore (Zhang et al., 2019), which measures semantic similarity using contextual embeddings from pre-trained BERT models.

BERTScore provides Precision, Recall, and F1 values by comparing token embeddings, offering

a more robust evaluation of semantic alignment. After evaluating the system across a spectrum of questions of varying complexity, we obtained the F1 scores shown in Table 2.

Query Type	Simple	Moderate	Complex
F1 score	0.82	0.77	0.71

Table 2: F1 scores for different query complexities

4.2 Human Evaluation

A panel of Nepali lawyers and law students performed a human review in addition to the computerized one to evaluate the system's practicality. Five-point ratings were assigned to responses based on five important metrics: Faithfulness: Consistency with legal texts; Relevance: Alignment with the query; Logical Correctness: Soundness of reasoning; Completeness: Coverage of necessary details; Interpretability: Clarity for users.

Metric	Sample Query Rating
Faithfulness	4.5/5
Relevance	5/5
Logical Correctness	4/5
Completeness	4/5
Interpretability	4/5

Table 3: Evaluation ratings for a sample query across various metrics

4.3 Discussion

The outcomes demonstrate how well the algorithm handles simple legal questions. Strong performance in providing precise and query-aligned answers for factual questions is indicated by the BERTScore of 0.82 for basic queries and a flawless Relevance score of 5.0/5 in human evaluation. The RAG framework's ability to anchor responses in legal texts and guarantee factual consistency is further demonstrated by the high Faithfulness score of 4.5/5.However, when query complexity increases, a noticeable drop in performance is observed. For complicated queries, the BERTScore falls to 0.71, while the human ratings for Completeness and Logical Correctness decrease to 4.0/5. These findings suggest that the system struggles to synthesize and explain nuanced or ambiguous legal concepts, particularly in complex queries. This highlights the problem that previous legal NLP studies have shown (Chalkidis et al., 2020) (Ryu et al., 2023), which is that automated measures frequently fall short of capturing complex legal thinking, particularly for ambiguous laws. Although the system is a useful tool for retrieving basic legal information, it has to be improved in order to handle more in-depth legal analysis.

5 Conclusion

This work represents a significant advancement in legal AI for low-resource languages, as demonstrated through our Nepali legal assistant system. We have created a tool that provides precise and straightforward responses to legal questions in Nepali, based on reliable sources, by combining a RAG framework with a refined LLM. This work democratizes legal information by offering a workable way to improve legal accessibility in Nepal. Our system gives people a basic grasp of their legal rights and duties, but it does not take the place of expert legal assistance, particularly in complicated instances.

6 Future Work

Future enhancements will focus on increasing the system's robustness and broadening its impact. Key areas include implementing mechanisms for automatic updates to the legal knowledge base to incorporate new laws and rulings, potentially through direct integration with official governmental databases, subject to privacy and technical constraints. Incorporating multimodal inputs like speech queries and extending language support to regional Nepali languages are also important directions aimed at creating a more accessible AI-driven platform for legal information in Nepal and potentially other low-resource contexts.

7 Limitations

The effectiveness of RAG retrieval may be limited by the absence of embedding models that are especially trained for the Nepali legal area. Although the 10,000-pair Q/A dataset is of good quality, its size may limit generalization by failing to completely capture rare or extremely complicated legal issues. Peak model performance was probably affected by efficiency-focused fine-tuning constraints (QLoRA, trained on only two epochs) as opposed to more thorough training. Reliability in intricate legal reasoning may also be impacted by the model's difficulties with ambiguous or interpretable legal texts.

8 Acknowledgements

We would like to express our sincere gratitude to students of **Kathmandu University School of Law** for helping us by performing human assessment. We would also like to thank the reviewers for their feedback and comments.

References

Benjamin Alarie, Anthony Niblett, and Albert H. Yoon. 2016. Regulation by machine. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*.

Bal Krishna Bal. 2004. Structure of nepali grammar. Technical report, Madan Puraskar Pustakalaya.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chroma Contributors. 2023. Chroma: The ai-native open-source embedding database. Https://www.trychroma.com.

Tim Dettmers and et al. 2023. Qlora: Efficient fine-tuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Sharad Duwal, Suraj Prasai, and Suresh Manandhar. 2024. Domain-adaptative continual learning for low-resource tasks: Evaluation on nepali. *Preprint*, arXiv:2412.13860.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Eric Han and Timo Schick. 2023. Unsloth: Fastest llm finetuning library. Https://github.com/unslothai/unsloth.

Edward J. Hu and et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Preprint*, arXiv:2005.11401.

Lightning AI. 2024. Lightning AI Studio. https://lightning.ai/lightning-ai-platform/.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*.

Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for Nepali text classification. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 106–111, Marseille, France. European Language Resources Association.

Meta AI. 2024. Llama 3.2 models are coming: Revolutionizing on-device and multimodal ai with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Prabin Paudel, Supriya Khadka, Ranju G.C., and Rahul Shah. 2024. Optimizing Nepali PDF extraction: A comparative study of parser and OCR technologies. *Preprint*, arXiv:2407.04577.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: A case study on indian law. In *Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023)*.

Shabdapurush Poudel, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional English-Nepali machine translation(MT) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 53–58, Torino, Italia. ELRA and ICCL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP*.

Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based evaluation for LLMs: A case study in Korean legal QA. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 132–137, Singapore. Association for Computational Linguistics.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. NepBERTa: Nepali language model trained in a large corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 273–284, Online only. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wayne Xin Zhao, Kun Zhou, Junran Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jingsen Jiang, Rui Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.