End-to-End Nepali Legal Question Answering: Building Linguistic Resources and Evaluating LLMs

Anjan Parajuli
Department Of Computer Science
Tribhuvan University
Balkhu,Kirtipur
anjan.765421@prnc.tu.edu.np

***Abstract*—**Legal knowledge is essential for informed decision-making; however, it remains largely inaccessible to the general public due to the high cost of legal services and the complexity of legal terminology. While many researchers have attempted to address this challenge by developing automated systems using large language models (LLMs) in high-resource languages, the Nepali language remains underexplored, primarily due to the limited availability of domain-specific legal resources. Therefore, this work aims to bridge this gap by developing key legal resources and evaluating LLM performance in the Nepali legal domain. We created the largest validated Nepali legal lexicon containing 48,117 unique tokens and a high-quality legal corpus of 59.6 MB, cleaned and validated from 673 official federal-level legal documents, including all active laws, regulations, and the Constitution. Furthermore, we analyzed 105 real-world legal questions spanning 10 categories under both Retrieval-Augmented Generation **(**RAG) and non-RAG settings, using a novel four-dimensional evaluation framework that assesses factual correctness, legal completeness, citation correctness, and reasoning ability. Experimental results show that RAG-based setups with semantic chunking outperform non-RAG configurations, improving overall rating from to 3.67 to 3.93. These findings highlight the importance of retrieval grounding and structured corpus representation in enhancing answer quality for low-resource legal question answering. To the best of our knowledge, this is the first comprehensive study of Legal Question Answering in the Nepali legal domain.

Keywords—Nepali Legal Question Answering, Legal Corpus, Legal Dictionary, Large Language Models, Retrieval-Augmented Generation

I. INTRODUCTION

According to the World Justice Project's Access to Justice survey (2019) [1], 84% of Nepalese faced at least one legal problem within two years, yet only 42% obtained expert assistance. Most relied on friends and family (83%), while just 11% consulted a lawyer and 1% accessed government legal aid. Although recent efforts—such as UNDP's legal aid for 50,000 marginalized individuals [2] and the Ministry of Law's support for 35,529 people in 2024 —have improved outreach, significant gaps remain. These findings underscore the urgent need for accessible, reliable tools to assist citizens in navigating legal issues.

Recent advancements in natural language processing (NLP) have enabled large language models (LLMs) to understand and generate human-readable text, creating new opportunities to improve information accessibility and address the challenges in legal knowledge dissemination. Despite substantial progress in legal question answering (QA) for languages such as English, Chinese, and Hindi, similar developments for Nepali remain extremely limited. Existing LLMs have not been systematically evaluated on Nepali legal questions, primarily due to the lack of domain-specific corpora, standardized legal terminology, and benchmark datasets. A few notable efforts, such as *NepKanun* [3], exist but remain constrained by limited and unstructured legal resources.

Motivated by these challenges and inspired by advancements in legal question answering for other languages, this work proposes an end-to-end framework for Nepali Legal Question Answering (LQA) leveraging Retrieval-Augmented Generation (RAG). The proposed approach integrates three core components:

- construction of a high-quality Nepali legal corpus by collecting and cleaning 673 official legal documents,
- development of a validated Nepali legal dictionary comprising 48,117 unique words to standardize terminology and enable effective text preprocessing,
- implementation of semantic chunking and retrieval of relevant legal passages to support precise and contextually grounded answer generation.

To encode the documents and their chunks, the Gemini Embedding 01 model was used, and the resulting embeddings were stored in a FAISS vector database for efficient similarity search. During the retrieval process, a two-stage strategy was employed: first, identifying the most relevant documents, and subsequently retrieving the top five most relevant chunks from each selected document. Model performance was evaluated on 105 real-world legal questions spanning 10 categories under both RAG and non-RAG settings, using four evaluation criteria—factual correctness, legal completeness, citation correctness, and reasoning ability.

Our experimental results demonstrate that integrating RAG significantly improves answer quality, achieving 97% accuracy on expert-evaluated responses. The key contributions of this work are:

- the creation of the largest high-quality Nepali legal corpus (61 MB) suitable for NLP applications,
- development of a validated Nepali legal dictionary enabling standardized preprocessing and semantic understanding,
- the analysis of a real-world legal question answering combining semantic chunking with RAG for effective retrieval and answer generation,

- the establishment of a benchmark evaluation on real-world Nepali legal questions across factual, legal, citation, and reasoning criteria and,
- the proposal of a two-stage retrieval strategy that first identifies the most relevant documents based on semantic similarity and subsequently retrieves the most relevant chunks from those documents. This hierarchical approach improves retrieval precision and reduces contextual noise during generation.

Collectively, these contributions provide a foundation for future research in Nepali legal NLP and demonstrate that retrieval-augmented reasoning can effectively bridge gaps in LLM performance for low-resource, domain-specific languages.

II. LITERATURE REVIEW

Legal Question Answering has been extensively explored across diverse languages and legal systems, employing a range of retrieval-based, generative, and hybrid techniques. To provide a structured overview, we categorize prior research into two broad groups based on language scope:

A. Legal QA research in non-Nepali languages

Legal Question Answering (Legal QA) has been explored in a limited number of languages, with English leading the research, followed by Vietnamese, Thai, Chinese, Japanese, Indian, German, and French. Representative studies include English works such as CBR-RAG [3] and DiscolQA [4], Vietnamese QA research with Question Analysis for Vietnamese Legal QA [5], and emerging Thai studies using the NitiBench dataset [6]. Japanese and French research has explored cross-lingual and multilingual transfer for legal QA, utilizing COLIEE [7] for Japanese - English legal text tasks and LLeQA [8], a French-language dataset, for French legal question answering.

Building on these datasets and representative studies, researchers have employed a variety of methodologies to address legal question answering. Retrieval-augmented generation (RAG) approaches, such as CBR-RAG [3], combine document retrieval with generative answer synthesis. Retrieve-and-read pipelines [8] first identify relevant legal texts and then extract or generate answers. Entailment-based methods [9] determine whether legal passages logically support a given question, often using transformer models. Neural attentive networks [10], FrameNet-based semantic parsing [11], and classical machine learning models such as SVM or deep CNNs [12] have also been explored for classification and reasoning tasks. Some studies follow a pretrain–finetune–rerank paradigm [13], leveraging large language models adapted to legal corpora. More recent approaches incorporate graph-based retrieval and generation (GraphRAG) [14] [15] to enhance reasoning over interconnected legal documents. Together, these methodologies highlight the diversity of strategies applied across English, Vietnamese, Thai, Chinese, Japanese, and French legal QA research.

B. Legal QA research in the Nepali language

While several languages have received considerable attention in legal QA, research in Nepali remains limited. Only a few notable studies have addressed general Nepali QA, including work using BERT models (extractive QA) [16] and Nepali GPT (open-ended QA) [17]. In the legal domain, Enhanced Retrieval for QA System Tailored for Nepali Legal Documents Focusing on PSC Examinations Using GPT-4 [18] and Nep Kanun (RAG-based) [19] stand out as the only dedicated Nepali legal QA studies to date. This highlights a significant gap and underscores the need for further research in Nepali legal question answering systems.

III. METHODOLOGY

Creating cleaned and quality legal corpus and analyzing them on LLM consisted of multiple stages which are outlined below:

A. Data Collection

For this study, two types of data were collected: legal documents and real-world legal questions. .

1) Legal Corpus: A total of 673 official Nepali legal documents—comprising 1 Constitution, 353 Acts, and 319 Regulations—were obtained in PDF format from the Nepal Law Commission's official repository [20]. Many of these PDFs were image-based, so Tesseract OCR was employed to convert them into machine-readable text, yielding approximately 59.6 MB of raw textual data. The extracted text contained OCR-induced artifacts, formatting inconsistencies, and typographical errors, as illustrated in Table I.

TABLE I. TYPES OF CORPUS NOISE

Text	Category
१५४//४/.|0//0011110189101.00४.1100	Noisy line
१८ गणतन्त्र सुदृढीकरण तथा केही नेपाल कानून संशोधन गर्ने ऐन, २०६६ द्वारा झिकिएको ।	Footer Noise
ÃÂ©	D*isallowed characters*
•·●◆■▪◦Ã□Â	*Jargon Characters*

2) Real-World Questions for RAG Evaluation: To evaluate the retrieval-augmented generation (RAG) framework, 100 real-world legal questions were manually curated. Ten questions were selected from each of 10 different legal categories, summarized in Table II. These questions were sourced from publicly available Nepalese law discussion groups on Facebook and the MeroAdda.com website [21]. The questions were originally in English, Romanized Nepali, and Nepali scripts, and all were normalized to Nepali to ensure consistency. This dataset served as a benchmark for assessing model performance across diverse legal topics.

TABLE II. QUESTION CATEGORIES

Category	No. of Questions
Constitutional and Administrative	11
Criminal and Justice	10
Civil and Private Dispute	11
Commercial Business and Economy	10
Labor and Employment	10
Taxation and Public Finance	10
Property and Land	10
Natural Resources	9

Social Welfare, Health and Education	11
Technology, Media and Intellectual Rights	13

B. Data Preprocessing

To preprocess data, a two-level text cleaning pipeline was developed to systematically remove noise, implemented in Python using regular expressions, Unicode normalization, and rule-based heuristics.

1) Character-level Cleaning :

- Non-Devanagari and invalid Unicode symbols were filtered using Unicode range-based validation.
- Only characters from the Devanagari, Devanagari Extended, and General Punctuation blocks were retained.

2) Text-level Cleaning

- Unicode normalization (NFC) was applied to standardize character encoding.
- Footer and header noise (e.g., page numbers, document titles) were removed using regex and repetition-based heuristics.
- Numeric pattern filtering eliminated irregular sequences of digits, mixed alphanumeric artifacts, and repeated parenthesized numbers.
- High-frequency noisy lines (appearing four or more times) were removed to reduce repetition.
- Whitespace normalization ensured consistent formatting across all documents.

After cleaning, the corpus was reduced to 56.5 MB, resulting in a uniform and high-quality textual dataset suitable for downstream NLP tasks.

C. Token Validation and Correction

From the cleaned legal corpus, a dictionary of 74,015 unique tokens was initially created by splitting text on spaces and removing disallowed characters. This dictionary was cross-checked against two validated Nepali word lists: NepWords [22] and POS Words [23]. Tokens present in these reference lists were marked as known and required no further validation. Tokens not found in the reference lists were subjected to automatic correction using Gemini, generating candidate corrected forms (Fig. 1.). To quantify similarity between original and corrected tokens, the Levenstein distance (LV distance) [24] was computed. Based on this metric, tokens were categorized for further validation into two groups:

1) Distance ≤ 3: Minor differences, requiring minimal verification.

2) Distance > 3: Substantial differences, requiring manual review.

This systematic process ensured that known tokens were preserved, while previously unknown or noisy tokens were corrected and validated. By combining the corrected tokens generated through Gemini with the pre-validated entries, we produced a final dictionary comprising 48,117 unique tokens, representing the largest high-quality Nepali legal lexicon developed to date for preprocessing and downstream NLP tasks. Moreover, the original 74,015-token lexicon was used to clean the corpus by replacing tokens if it was corrected from Gemini. The resulting corpus constitutes the largest Nepali legal text collection created so far.


```
{
        "given": "निकायह,
        "is_corrected": true,
        "corrected": ["निकायहरूले"]
}
```


Fig. 1. Corrected output from Gemini

D. Semantic Chunking

To enable efficient and contextually accurate retrieval, each legal document was segmented into semantically meaningful units through a two-stage chunking pipeline.

1) Document Classification: Each document was first analyzed to determine whether it contained परिच्छेद (grouped sections of legal provisions). This classification guided the subsequent chunking strategy, ensuring structural consistency across varying document formats.

2) Chunking: Customized regular expressions were designed for documents with and without परिच्छेद. The text was segmented into distinct semantic components such as introduction, धारा (articles), अनुसूची (appendices), and द्रष्टव्य (end statements). To balance granularity and computational efficiency, each chunk was limited to a maximum of 2000 characters. Oversized chunks were further split using indexed markers in their metadata, while smaller adjacent chunks were merged to maintain coherence. Each chunk was annotated with metadata including the document name, धारा (article) name, and परिच्छेद (section) identifier, as illustrated in Fig. 2.


```
{"idx": 120,
"doc": {"meta":
        {"content_type":"dhaara","doc_name":"राजस्व चुहावट (अनुसन्धान तथा नियन्त्रण) ऐन, २०५२",
"paricched_name":"अनुसन्धान तथा तहकिकात",
"paricched_number": 3,
"dhaara_name": ["थुनामा राख्न सकिने", "३झेख. कम्प्युटरबाट प्रशोधित अभिलेख प्रमाणको रूपमा लिन सक्ने", "१३ग. वेबमा आधारित अनलाइन दुवानी साधन अनुगमन प्रणाली लागू गर्ने"], "dhaara_number": [17, 18, 19]},
 "text": ".....",
“embedding: […]
```


Fig. 2. Chunk Format

E. Embedding

Following the chunking process, vector embeddings were generated using the Gemini Embedding-01 model to capture semantic representations of the legal text. Two hierarchical levels of embeddings were created to support both broad and fine-grained retrieval:

1) Document Level Embeddings: Each document title was embedded and indexed in a FAISS vector store, facilitating high-level semantic retrieval across the entire legal corpus.

2) Chunk Level Embeddings: Each chunk's textual content was independently embedded and stored in a dedicated FAISS index, enabling fine-grained retrieval for RAG pipeline.

F. Retrieval and Generation Strategy

To optimize both relevance and response efficiency, a hierarchical retrieval strategy was implemented:

1) Document-Level Retrieval: For each query, the top five most relevant documents were identified using vector similarity search over the document-level FAISS database.

2) Chunk-Level Retrieval: Within each retrieved document, the top five most relevant chunks were selected based on vector similarity from the chunk-level FAISS database.

These selected chunks were then provided as input to the RAG model, ensuring that generated answers were contextually precise and grounded in the most pertinent legal text. The Gemini Embedding-01 model was employed throughout the retrieval process. The complete workflow is illustrated in Fig. 3.

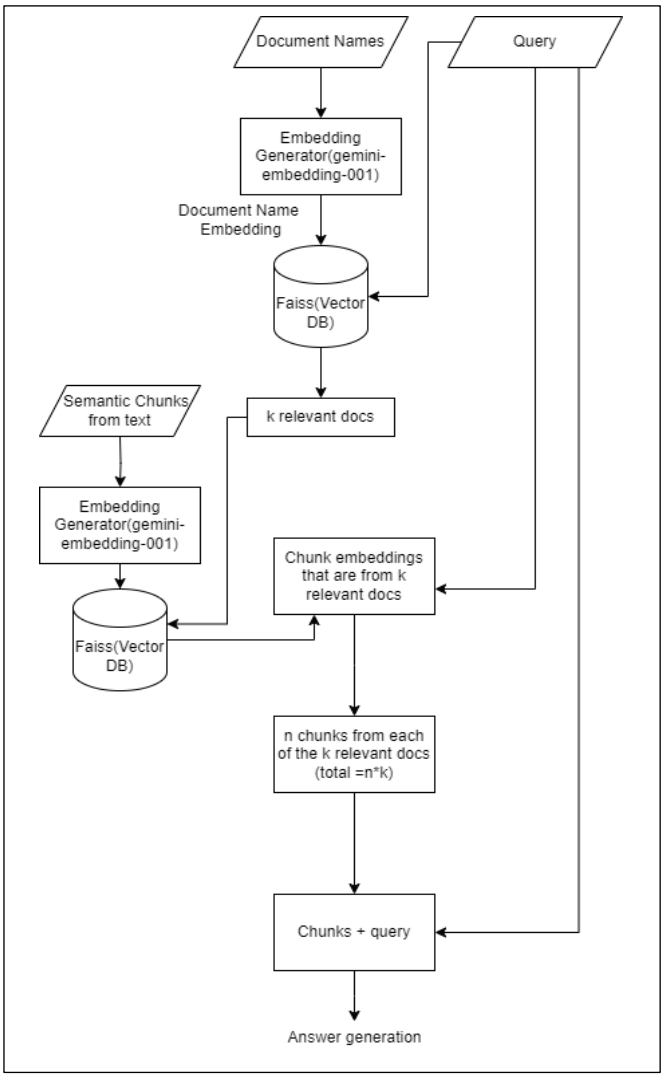

Fig. 3. System Diagram for the experiment

IV. EVALUATION

A. Evaluating Dictionary and Corpus

To assess the quality of the constructed dictionary, a total of 74,015 tokens were analyzed. Among them, 14,548 tokens were directly recognized as valid entries after passing through Nepwords [32] and POS words [33]. The remaining 59,467 tokens were grouped based on their Levenhtein (LV) distance, which measures the minimum number of edits required to transform one token into another. Three groups were formed:

- Group 1: LV distance ≤ 3
- Group 2: LV distance > 3
- Group 3: LV distance = 0

From each group, a random sample of 500 tokens was manually verified. The accuracy for each group was calculated using the formula:

$$Accuracy = \frac{Total\ number\ of\ correct\ tokens}{Total\ number\ of\ tokens}$$

After incorporating all manually verified and corrected mappings, a total of 48,117 validated tokens were obtained (including the initial 14,548 valid entries). Consequently, the resulting dictionary represents the largest validated Nepali legal lexicon developed to date, suitable for subsequent text preprocessing and analysis.

Furthermore, to evaluate the quality of the compiled corpus, dictionary coverage was assessed by calculating the proportion of tokens present in the corpus that matched entries in the validated dictionary. The corpus accuracy was computed using:

$$Accuracy = \frac{Tokens\ recognized\ by\ dictionary}{Total\ tokens\ in\ corpus}$$

B. Evaluating RAG answers

To assess the quality of generated answers, a four-factor evaluation framework was employed. Each response was evaluated across four dimensions, with one point assigned per criterion, yielding a maximum total score of 4. The factors were defined as follows:

- Factual Correctness – Assesses whether the answer is consistent with the factual ground truth.
- Legal Completeness – Evaluates the extent to which the response covers all legally relevant aspects from the provided context.
- Citation Correctness – Checks the accuracy and appropriateness of cited legal provisions.
- Legal Reasoning – Examines the coherence and interpretative validity of the argument presented.

The final score represents the sum of all criterion scores (0–4 range), providing a balanced measure of factual accuracy, legal adequacy, and interpretative soundness.

V. RESULTS AND DISCUSSION

We evaluated the quality of the dictionary, legal corpus, and retrieval-augmented generation (RAG) answers using the metrics described in the previous section. The quantitative results are summarized in Tables III, and IV respectively.

TABLE III. DICTIONARY ACCURACY

Group	Accuracy
(distance <=3)	97.8%

(distance>3)	96.4%
(distance=0)	99.4%
Average	97.8%

TABLE IV. RAG ANSWERS EVALUATION

Test type	Average Rating	Percentage Accuracy
Without RAG	3.67	91.75%
With RAG	3.93	98.25%

A. Dictionary Evaluation

As shown in Table III, the evaluation of three token groups (Group 1, Group 2, and Group 3), from 59467 Gemini corrected tokens, yielded accuracies of 97.8%, 96.4%, and 99.4%, respectively, with an average accuracy of 97.8%. This high validation rate demonstrates that the constructed dictionary is both accurate and reliable, and therefore suitable for downstream text preprocessing and linguistic normalization tasks in Nepali NLP pipelines.

B. Corpus Evaluation

The compiled legal corpus achieved a dictionary coverage rate of 99.7%, indicating that nearly all tokens in the corpus correspond to valid entries in the standardized dictionary. This high overlap confirms that the corpus is lexically consistent and cleaned of noise, making it well-suited for domain-adaptive pretraining of large language models (LLMs).

C. RAG Answers Evaluation

For evaluating generated answers, we used the FLCL metric, which measures Factual Correctness, Legal Completeness, Citation Correctness, and Legal Reasoning. The RAG responses were produced using Gemini, as it exhibited strong Nepali comprehension and QA ability. As shown in Table IV, the system achieved an overall accuracy of 98.25% (computed from the average rating) using RAG across the FLCL components, with each factor receiving high ratings from a legal expert evaluator. The analysis reveals that non-RAG-based responses demonstrated limitations in delivering recent information and failed to include adequate citations.

These results collectively indicate that both the linguistic resources (dictionary and corpus) and the QA system (Gemini) demonstrate a high level of linguistic and legal fidelity, reinforcing the robustness of the pipeline for Nepali legal question answering.

VI. CONCLUSION AND FUTURE WORKS

This study presented a comprehensive framework for Nepali Legal Question Answering, featuring the creation of a validated legal dictionary, a cleaned domain-specific corpus, and a RAG-based evaluation pipeline. The dictionary achieved 97.8% accuracy, while the corpus showed 99.7% lexical coverage, confirming their quality for NLP preprocessing and domain-adaptive pretraining. Additionally, the RAG model, evaluated using the FLCL metric, attained 98.25% overall accuracy, highlighting strong factual and legal consistency. Future work will focus on expanding the dataset with case laws, fine-tuning domain-specific LLMs, establishing a standard Nepali Legal QA benchmark, and developing an interactive legal assistant to enhance public access to legal information.

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to Prof. Dr. Thakur Prasad Upadhyaya, Teacher Devi Lal Timilsina, and Assoc. Professor Chudamani Subedi, for their invaluable guidance, encouragement, and insightful feedback throughout the course of this research. Their mentorship and expertise were instrumental in shaping the direction of this study and ensuring its academic rigor. The author also extends appreciation to the legal experts who contributed to the manual evaluation process. Additionally, the author acknowledges the Nepal Law Commission and MeroAdda.com, from which the legal data were manually collected, providing essential resources that made this study possible.

ETHICAL CONSIDERATIONS

All legal documents used in this study were collected from publicly accessible sources, including the Nepal Law Commission and MeroAdda.com, ensuring compliance with copyright and intellectual property laws. No confidential or private records were accessed. Manual data validation and evaluation were conducted responsibly, with legal experts participating voluntarily. Generated RAG outputs were used solely for research purposes, and all results were handled carefully to prevent misinterpretation. This approach ensured that the study-maintained transparency, legality, and integrity in handling legal data.

REFERENCES

[1] S. Consultant, 2017. [Online]. Available: https://worldjusticeproject.org/sites/default/files/documents/Access-to-Justice-2019-Nepal.pdf.

[2] UNDP, "Annual Report 2023 Asia Nepal," [Online]. Available: https://rolhr.undp.org/annualreport/2023/asia-pacific/nepal.html .

[3] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji and B. Fleisch, "CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering," in *International Conference on Case-Based Reasoning*, Cham, 2024.

[4] F. Sovrano, M. Palmirani and S. Sapienza, "DiscoLQA: zero-shot discourse-based legal question answering on European Legislation," in *Artificial Intelligence and Law*, 2025.

[5] L. T. N. C. T. H. N. T. a. T. M. P. N. X. Bach, "Question analysis for Vietnamese legal question answering," in *9th International Conference on Knowledge and Systems Engineering (KSE)*, Hue, 2017 .

[6] P. P. P. C. C. T. P. N. T. P. K. &. N. S. Akarajaradwong, "Nitibench: A comprehensive study of llm framework capabilities for thai legal question answering.," in *arXiv preprint arXiv:2502.10868*, 2025.

[7] Y. Kano, M. Y. Kim, R. Goebel and K. Satoh, "Overview of COLIEE 2017," in *COLIEE@ ICAIL*, 2017.

[8] A. v. D. G. &. S. G. Louis, "Interpretable long-form legal question answering with retrieval-augmented large language models," in *In Proceedings of the AAAI Conference on Artificial Intelligence* , 2024.

[9] C. P. N. T. T. D. N. A. T. T. P. A. D. a. L.-M. N. Nguyen, "Captain at coliee 2023: Efficient methods for legal information retrieval and entailment tasks.," in *arXiv preprint arXiv:2401.03551*, 2024.

[10] A. K. D. S. M. S. H. &. M. Y. Morimoto, " Legal Question Answering System using Neural Attention," in *COLIEE@ ICAIL*, 2017.

[11] R. H. R. &. K. Y. Taniguchi, "Legal question answering system using framenet," in *In JSAI international symposium on artificial intelligence*, Cham, 2018.

[12] P. K. N. H. T. T. C. X. N. M. T. &. N. M. L. Do, "Legal question answering using ranking SVM and deep convolutional neural network," in *arXiv preprint arXiv:1703.05320.* , 2017.

[13] S. C. H. &. Y. M. Ni, "Pre-training, Fine-tuning and Re-ranking: A Three-Stage Framework for Legal Question Answering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.

[14] H. Zhai, "Law GraphRAG: An Advanced Legal Question-Answering System," in *International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)*, 2025.

[15] C. G. S. R. A. C. D. C. &. C. M. C. Crăciun, "GRAF: Graph Retrieval Augmented by Facts for Romanian Legal Multi-Choice Question Answering.," in *arXiv preprint arXiv:2412.04119.* , 2024.

[16] U. T. S. T. H. N. &. U. M. Thapa, "Nepali Question Answering System from Multilingual BERT Model and," in *Proceedings of 15th IOE Graduate Conference*, 2024.

[17] S. S. A. S. S. B. S. T. S. &. P. S. Pudasaini, "NepaliGPT: A Generative Language Model for the Nepali Language," in *arXiv preprint arXiv:2506.16399.* , 2025.

[18] A. L. D. S. B. Nabin Bhusal, "Enhanced retrieval for QA system tailored for Nepali legal documents focusing on PSC exams using GPT-4 and RAG framework," in *IOE Proceedings*, 2025.

[19] "NepKanun: A RAG-Based Nepali Legal Assistant," in *ACL*, 2025.

[20] L. C. Nepal, "pages/alphabetical-index-of-acts/," [Online]. Available: https://lawcommission.gov.np.

[21] [Online]. Available: https://meroadda.com/.

[22] Q. N. Van. [Online]. Available: https://github.com/quanap5/13Nov2018/blob/master/langdata-master/langdata-master/nep/nep.wordlist .

[23] S. G. M. &. B. B. Timilsina, "Nepberta: Nepali language model trained in a large corpus.," in *Association for Computational Linguistics (ACL)*, 2022.

[24] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady," 1966.