Advancing Legal Accessibility in Bangladesh through AI-Powered Assistance and Natural Language Interfaces

by

Tanzir Hossain
20301154
Md. Shakil Anawar
20301162
Ar-Rafi Islam
20301164
Md. Adnan Karim
20301157

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing a degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tanzir Hossain 20301154	Md. Adnan Karim 20301157
Md. Shakil Anawar 20301162	Ar-Rafi Islam 20301164

Approval

The thesis titled "Advancing Legal Accessibility in Bangladesh through AI-Powered Assistance and Natural Language Interfaces" submitted by

1. Tanzir Hossain (20301154)
2. Md. Shakil Anawar (20301162)
3. Ar-Rafi Islam (20301164)
4. Md. Adnan Karim (20301157)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 17, 2024.

Examining Committee:
Supervisor:
(Member)

Dr. Farig Yousuf Sadeque
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson & Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

In a country where laws and regulations are as difficult to understand as the cases they govern, the path to justice is anything but straightforward. For many in Bangladesh, the legal system's complexity acts as a gatekeeper because most of the people can not simply comprehend its complex language. Our research presents a Retrieval Augmented Generation (RAG) system that integrates large transformer-based language models like Mistral-7B, deCILM-7B, LLaMA 3.1-8B, and others with a specialized legal retrieval module containing Bangladeshi laws and regulations. This allows the system to interpret user inquiries using natural language, retrieve relevant legal information from the corpus, and generate detailed responses grounded in and applying the country's legal code - complete with applicable citations. At the heart of it all is an embedded Retrieval Augmented Generative system that combines neural automation and strong encoder-decoder transformers for generating legal documents and classifications quickly, ensuring legal accuracy. And the human-in-the-loop verification process to review and modify output allows users to get even more accurate results. This can massively improve general people's understanding of legal matters as well as it can work as a lawyer's helping hand too.

Keywords: Retrieval-Augmented Generation (RAG); Large Transformer-Based Language Models; (Mistral, deCILM-7B, LLaMA 3.1-8B); Natural Language Processing (NLP); Human-in-the-Loop Validation; LangChain Framework; Machine Learning, Artificial Intelligence.

Table of Contents

List of Figures

Nomenclature

The next list describes several symbols & abbreviations that will be later used within the body of the document

LLM	Large Language Model
RAG	Retrieval-Augmented Generation
BanglaNMT	Bangla Neural Machine Translation

Chapter 1

Introduction

Navigating the legal landscape in Bangladesh can feel like deciphering an elaborate code without a key. A gap between the complex legal framework and the general people's understanding of legal concepts remains to date [3], [13]. Being complex and difficult to understand is one thing however, when it comes to legal procedure the burden for the mass people feels even heavier [12]. Legal procedures in Bangladesh are often lengthy, costly, scarce, and filled with complicated languages making the navigation procedure through the legal system impossible for many. Lack of legal knowledge, shortage of lawyers, absence of legal resources, and high costs of legal services all of these together combined make the problem even more exacerbated. Resulting in people not getting any legal help they need. Not only does this problem hinder people's ability to defend their rights but also undermines access to justice. The barriers that keep us away from the desired legal aid are rooted in several factors. Firstly, the legal system in Bangladesh is highly specialized, secondly, legal texts are often being written in a language that is hardly comprehensible for a non-expert. In addition, the unavailability of lawyers in rural [1], the legal representation being highly costly, and the protracted nature of legal proceedings all together create such a system that often becomes out of reach for ordinary citizens[2]. There are some legal aid programs out there taking the initiative to provide legal assistance to the mass people. However, lack of resources, reaching the entire population becomes a big challenge resulting in a significant gap in service provision.

The traditional ways to solve such a problem or we can say addressing the accessibility of legal aid have heavily relied on legal aid clinics, non-governmental organizations, and governmental programs. These aid systems try to provide legal services to underrepresented populations [4]. The fact can not be denied that the above-mentioned initiatives have tried to make some impact. However, they are often limited by the constraints of funding, geographical isolation, and insufficient number of trained legal professionals. On the other hand, the legal resources that are available online typically offer static information that lacks personalization. To be more specific, the online platforms lack to adapt to specific user needs, and have fewer options for scenario-based problems, making it less useful in practical legal contexts. Even in the era of AI, the AI-based legal system suffers from inaccuracies, hallucinations, and lack of domain-specific knowledge resulting in an unreliable legal assistant. With the rapid evolution of AI, we are witnessing the dawn of a new age. Where science fiction becomes a reality the challenges we are facing regarding legal aid can be tackled with an AI-driven legal aid system[17]. It not only will be

more accessible to the mass people but also be reliable in providing accurate legal information. Recent advancements in LLMs and NLP techniques have opened many promising pathways to solve our problem [36]. However, integrating these technologies in a manner that ensures high accuracy, relevance, and legal compliance remains an obstacle. Especially for the legal domain of Bangladesh, this sector is yet to be explored.

1.1 Research Problem

In this thesis we will try to develop an AI-driven legal assistance system for Bangladesh, integrating advanced large language models and legal retrieval to deliver precise, accessible legal information and case references to general people, law students, and professionals.

1.2 Research Objective

- We will try to develop a retrieval-augmented generation system integrating large language models with a Bangladeshi legal corpus for natural language legal querying and response generation with citations.
- Evaluating the effectiveness of transformer language models and neural retrieval in interpreting and applying Bangladeshi legal code to user inquiries, focusing on accessibility and precision.
- Implementing a human validation system for legal experts to review will refine AI-generated legal outputs, ensuring accuracy and compliance.
- We will try to create an AI legal assistance, democratizing access to legal knowledge for individuals facing barriers, students, professionals, and others across Bangladesh.
- LangChain's integration of language models and retrieval techniques for robust natural language legal querying specific to the Bangladeshi context will be a part of our investigation.
- Developing techniques for scraping, curating, and structuring a large-scale Bangladeshi legal corpus for efficient retrieval and RAG system integration.

1.3 Research Contribution

- Introducing the very first Bangladesh's AI-based legal assistant for legal queries in natural language and retrieving relevant legal documents.
- With no existing dataset, we manually created three legal datasets for this research from scratch. i.e. Laws of Bangladesh(Golden Dataset), DailyStar law dataset, and Judgement dataset from open source.
- We fine-tuned two large language models, LLaMA 3.1 8B Instruct and Mistral 7B, using unsupervised learning techniques on our unlabeled law dataset.

- We implemented strategies such as semantic text splitting to mitigate hallucination in AI responses, thereby enhancing the accuracy and reliability of the assistant.
- We introduced a new evaluation metric called Legal-BERTScore, a domain-adapted evaluation metric that combines an F1 score based on semantic similarity and a hallucination score to comprehensively assess the semantic correctness and factual reliability of AI-generated legal responses.

1.4 Research Structure

The remainder of this paper is structured as follows:
In Chapter 2, we provide a thorough literature review. In Chapter 3, we show our datasets with visualizations of each data. It also shows the steps of preprocessing and cleaning the data with techniques like BS4, Selenium, Regular Expression, Pandas Library, and the translation model. Chapter 4 defines the model we have used for this research, specifically Llama3.1-7B Instruct RAG, Mistral-7B RAG, DeciAi, mxbai-embed-large-v1, etc., and how they work. Chapter 5 concludes the Human-In-The-Loop evaluation process. Chapter 6 is the detailed qualitative and quantitative result analysis that includes public and expert evaluation scores, BERT scores, LegalBERT scores, and some comparative analysis. In Chapter 7 we discuss error analysis where we show what our system is doing wrong and why. Finally, Chapter 8 concludes what this paper is about and suggests some future work.

Chapter 2

Literature Review

In our pursuit of Advancing Legal Accessibility in Bangladesh through AI-Powered Assistance and Natural Language Interfaces, we dedicated our efforts to thoroughly examining research papers within the domains of LLM, NLP, and RAG. Our selection process was intensive, particularly emphasizing papers containing high citation counts and recent publication dates. We have organized these selected papers chronologically based on their publication date, enabling us to build a cohesive and current knowledge foundation for our research efforts.

The paper of Soha Khazaeli et al. [11] introduced a sophisticated legal QA(Question-Answering) system aimed at answering complex legal questions without limiting the predefined patterns or specific legal practice areas. This system is mainly designed to assist legal professionals by providing answers to both factoid (simple, specific) and non-factoid (open-ended, explanatory) questions, in simple words according to the need. The primary contribution of this study lies in the retrieval-based approach it has made. In this retrieval-based approach, the combination of sparse vector search using BM25 with dense vector techniques like Legal GloVe and Siamese Legal BERT embeddings is used. In addition, these methods help the system to capture both keyword based and semantic relationships in the legal texts. Resulting in significant improvement in the relevance of retrieved passages. For this study the dataset used has over 100 million passages from legal documents, primarily case law headnotes. These headnotes help the system to learn nuanced legal language patterns. Moving on, fine-tuning was performed here on more than 10,000 annotated legal question-answer pairs, supplemented with additional general domain data to enhance model generalization. The authors have used a two-stage retrieval and ranking approach. In this approach, initial retrieval selects relevant passages, and a BERT-based re-ranking model (fine-tuned on Legal BERT) improves answer precision by scoring the likelihood. Here the likelihood is the chances of a passage correctly answering the input question. Now coming to the performance, a test set of 100 questions showed that the best system configuration achieved a DCG@3 (Discounted Cumulative Gain) score of 7.75. This score came through a combination of BM25 and Siamese Legal BERT embeddings, demonstrating the efficacy of combining different retrieval methods and deep learning techniques. The reason behind using such techniques is that they provide accurate and contextual relevant legal answers in a commercial setting. Overall it can be said that this work marks a significant advancement in legal NLP(Natural Language Processing). It also addresses the

challenges of varied legal language and high stakes associated with legal decision making.

Cheol Ryu et al. in their paper [18] introduced a novel approach for the evaluation of LLM (Large Language Model) model generated texts in the legal domain by proposing Eval-RAG. It is a retrieval-augmented evaluation framework. The primary contribution of this study lies in addressing the limitations of traditional LLM evaluations by integrating relevant legal documents into the evaluation process. The reason behind this is that the evaluation often overlooks factual errors and lacks domain-specific knowledge. In the paper, the mentioned dataset comprises three types of publicly available legal documents. Those are legal question-answer pairs, legal provisions, and legal case precedents, amounting to 287 QA pairs, 84 legal provisions, and 240 cases. This dataset provided a robust foundation for evaluating legal QA tasks. The study in their methodology part involved retrieving relevant documents based on a given query using a retriever which eventually uses sentence embeddings for similarity matching and then evaluates the generated answer against those documents to verify its factual accuracy. Models like 'test-embedding-ada-022' and GPT-4 were employed for embedding generation, generating relevant question-answers. Also, Eval-RAG was used to enhance existing evaluation methods like FairEval and ChatEval through incorporating retrieved documents into the evaluation prompts. By combining Eval-RAG with FairEval on GPT-4 the research yielded the highest correlation with human evaluations. It achieved a Pearson correlation of 0.5923. The results clearly outperformed standard LLM-based evaluation approaches. In the field of NLP, this research contributes by improving the reliability of LLM-generated text evaluation, particularly in tasks that need precise legal knowledge and document-based reasoning.

The research [21] by Shamane Siriwardhana et al. addresses a novel approach to enhance RAG models' performance in domain-specific open-domain question answering (ODQA). The authors found that RAG primarily trained on Wikipedia-based knowledge, struggles to adapt to specialized domains like healthcare and news. After that, they introduced RAG-end2end as an extension that allows joint training of the retriever and generator components of RAG. This method not only ensures the model adapts more effectively to domain-specific contexts but also updates both the knowledge base and the retriever during training. In addition, the authors proposed an auxiliary training signal that injects domain-specific knowledge by reconstructing sentences from a knowledge base. This dual of joint retriever-generator training and auxiliary signals helped the study to achieve significant performance improvements. For the study, the authors used three distinct datasets: 1.COVID-19 research, News, and Conversations. They used the CORD-19 dataset for the COVID-19 domain. It helped to create a knowledge base of over 250,000 passages and generated synthetic QA pairs using a BART model. Also, the authors created datasets for the News and Conversations domains in a similar fashion. The experiments from the study show that RAG-end2end outperformed the original RAG significantly with improved matrices across all domains. To be more specific, in the Conversations domain, the RAG-end2end model achieved an EM (Exact Match) score of 25.95% and an F1 score of 37.96%, while in the News domain, the highest F1 score was 23.7% with the RAG-end2end-QA+R model. The study highlighted that for better domain adap-

tation fine tuning the retriever, alongside the question encode was crucial. With the use of the RAG-end2end-QA+R model the study achieved a Top-20 retrieval accuracy of 50.95%, its highest recorded accuracy was for the News domain. The highest accuracy was clearly a big jump from the original RAG model.

The research by Akari Asai et al. [14] introduces a novel framework called SELF-RAG to address factual inaccuracies in LLMs (Large Language Models) by integrating retrieval and self-reflection mechanisms. The authors' motivation behind this work is that while RAG (retrieval-augmented generation) improves factual grounding in language models by fetching relevant information from external sources, it often retrieves irrelevant or unnecessary passages, eventually reducing generation quality. This problem is tackled by SELF-RAG by allowing the model to dynamically decide when to retrieve information and critically assess both the retrieved passages and its own generated outputs using self-reflection tokens. After that the reflection token allows the model to determine the relevance of retrieved passages and evaluate its generated outputs for factuality and quality. After all these rigorous processes the model becomes more controllable during the interface. Furthermore, retrieval tokens are used by SELF-RAG to decide whether to retrieve external documents. Also, critique tokens were used to evaluate the relevance, support, and utility of the generated outputs. After that, this method went through several tasks to be tested with open-domain question answering (QA), reasoning, and fact verification. Coming to the performance, The proposed model outperformed models like ChatGPT with strong baselines. Also, it outperformed retrieval-augmented Llama2-chat across various benchmarks. In this study, PopQA and Trivia QA named datasets were used. On PopQA SELF-RAG got an accuracy of 55.8% and on TriviaQA it got 74.5%, which is also the highest accuracy. The importance of retrieval mechanisms and critique tokens to achieve better factual accuracy and citation precision was highlighted in this study. The significant enhancement in the performance of LLM in knowledge-intensive tasks is not ignorable.

The authors Ruihao Shui et al. [20] presented the capabilities of large language models (LLMs) in the domain of law, specifically for legal judgment prediction(LJP), in their study. Determining charges based on case facts was quite a challenge and essential task in a legal system according to the authors. They explored models like GPT-4, ChatGPT, Vicuna, BLOOMZ, and ChatGLM and compared them with simpler models or information retrieval (IR) systems in handling complex legal reasoning tasks. The motivation behind this work comes from the doubts surrounding the performance of LLMs. To be more specific, GPT-4 claims its proficiency in real-world legal tasks, despite its proficiency in standardized law exams. As a result, the authors implemented two work scenarios for evaluation, 1. LLMs independent operations, 2. LLMs collaboration with IR system. The second method will help to retrieve similar legal cases or legal candidates. For the evaluation part, a dataset named CAIL was used. The dataset includes Chinese legal cases. After that the models were tested in four settings: 1. zero-shot, 2. few-shot, 3. open questions, 4. Multiple-choice questions. Four of these settings allowed the authors to test how models benefit from having similar cases or label candidates included in their prompts. The results indicated that adding the domain-specific information through retrieval systems helped improve accuracy. Specifically, complex models

like GPT-4 and ChatGPT, with GPT-4 achieved the highest accuracy of 74.46% in few-shot multiple-choice questions. However, the authors have left us in a paradox as in some cases, the IR system itself outperformed the combination of LLM and IR, suggesting that LLMs did not fully leverage the additional information. A comprehensive evaluation framework that can be applied across different domains and the retrieval-augmented generation's potential to significantly improve legal AI systems is the main contribution of this paper. It also added that LLMs can better integrate retrieved knowledge.

The study [30] of Nishchal Prasad et al. addresses the challenge of legal judgment prediction for long, unstructured legal texts that sometimes exceed tens of thousands of words. The study also raised the issue of legal documents being highly complex not for only being lengthy but also for their inconsistent structure and specialized language. To tackle this issue, the authors developed a hierarchical classification framework named MESdc (Multi-stage Encoder-based Supervised with Clustering). This framework divides the documents into smaller parts, then extracts embeddings from the last four layers of a fine-tuned LLM (Large Language Model), and finally approximates the document structure using unsupervised clustering. All of these embedding and structure approximations are then processed through additional transformer encoder layers for classification. For the models, the study utilized LLMs like GPT-Neo and GPT-J and compared their performance both in isolation and within the hierarchical MESc framework. The datasets used in this research are ILDC (Indian Legal Document Corpus) and subsets of the LexGLUE dataset, which includes European and U.S. legal documents. The authors showed that the hierarchical MESc framework yielded a 2% improvement over previous state-of-the-art models by fine-tuning these LLMs and conducting extensive ablation studies. The GPT-J model provided the highest accuracy, especially when fine-tuned with longer input lengths (2048 tokens). It also highlighted the importance of processing longer sections of documents. In the field of legal NLP, this paper contributes significantly by demonstrating the potential of LLMs in handling large, unstructured texts and establishing new baselines for legal document classification.

In this paper [37], the authors have proposed a new method called ConvRAG which enhances Conversational Question Answering (CQA) using fine grained retrieval augmentation and self check mechanism. The authors take two challenges: i. better understanding of context based questions and ii. getting proper knowledge for open-domain queries. To solve this they have proposed a three component method: a question Refiner for better understanding, a Fine-Grained Retriever, and a Self-Check Response Generator to filter garbage info. They also introduced a new Chinese CQA dataset with extended features. They test their approach with existing systems like QReCC and TOPIOCQA and also their own dataset. They compared ConvRAG with ChatGPT, WebPCM, and Perplexity.ai. They tried various evaluation metrics such as BLEU, ROUGE, METEOR, and GPT-4. The results show that the ConvRAG works better than the baseline achieving a 10% increase in ROUGE-L scores on the test dataset. This approach also outperformed the GPT model and Perplexity.ai. This shows the effectiveness of this approach. In this paper, the highest reported accuracy is BLEU-1 on the test dataset using the full ConvRAG model.

This study [31] shows the method of improving current RAG pipelines in the context of answering financial documents. Shetty et al. show major limitations in current RAG pipelines. They focused mainly on enhancing text retrieval for LLMs in domain specific tasks. They show many difficulties with current RAG pipelines including uniform chunking, limitation of semantic search, inability to control complex reasoning, etc. for financial analysis. To address these they proposed some techniques like sophisticated chunking methods, query expansion using HyDE, metadata annotation, and fine tuning embedding models. The authors used the FinanceBench dataset which consists of 10,231 questions about publicly traded companies from 2015-2023. They used metric systems like BLEU and ROGUE-L. Though the paper doesn't disclose the results, it emphasizes the importance of effective retrieval. It highlights that the best generators can even give incorrect answers. They also contribute to improving the RAG system, particularly in finance.

In this paper [32], the authors aimed to highlight the limitations of current RAG approaches while solving long context based documents in question answering. This paper by Shi et al. shows a concept based RAG framework that works with Abstract Meaning Representation (AMR) to improve questions answering in context based problems. First, they address the problems with current RAG pipelines such as not being able to handle complex scenarios and including irrelevant information. They proposed AMR based algorithms that take a lot of information and compress it into a small informative chunk. They used two datasets: PopQA and EntityQuestions. They modified these datasets to focus on mainly scenarios. They compared their work against various compression techniques such as keyword extraction, summarization, etc, and LLMs like GPT-Neo, OPT, BLOOM, and LLAMA-2. Their suggested framework outperformed all of them. The authors reported significant improvements over baseline methods in solving context based problems. The LLAMA-2-13b-chat-hf- model combined with RAG achieved the high performance improvement. On the PopQA dataset, it performed 71.43% better with RAG in long context scenarios. The authors in this paper demonstrated AMR to combine with RAG and developed an AMR based algorithm. They also highlighted the approach's ability to compress a context by over 60% while not changing the core information.

This paper [28] shows the performance of LLMs and traditional legal contract reviewers, specifically junior lawyers and LPOs. The authors aimed to show whether LLMs can outperform humans in terms of accuracy, speed, and cost efficiency. Though AI has shown great potential to solve problems in different fields, still there is a help in retrieving legal information correctly and factually in contracts. To address this gap, this paper focuses on three questions : (1) Do LLMs outperform Junior Lawyers and LPOs in determining and locating legal issues in contracts? (2) Can LLMs review contracts faster than Junior Lawyers and LPOs? and (3) Can LLMs review contracts cheaper than Junior Lawyers and LPOs? They utilized a dataset of contracts from real world agreements. A senior lawyer established a ground benchmark and they compared the results with this benchmark. Various LLMs like GPT-4, GPT-3.5, Claude2.0, and Palm2 were competing. The authors used metrics like precision, Recall, and F-score. They also checked efficiency and cost effectiveness. They found that GPT4-1106 outperformed humans. GPT4-1106 and LPOs both scored the highest F-score of 0.87. Notably, the LLMs were much

faster and cost efficient. The Palm2 took an average of 0.73 minutes compared to Junior Lawyers 56.17 minutes and 201 minutes for LPOs. It was also cost effective. The Claude 2.0 cost $0.02 per document whereas Junior Lawyers cost $74.26. The study clearly suggests that the LLMs are performing better in reviewing contracts than traditional lawyers and LPOs. It is also fast and cost-effective. It truly shows the potential of AI in legal work frames.

In this paper [16], the authors have proposed a framework named Prompt, Generate, Train (PGT) to develop a question answering model for open book questions answering over exclusive document collection. The authors have attempted to address the challenge of adapting retrieval augmented generation (RAG) models to specific domains using supervised fine-tuning and reinforcement learning with synthetic feedback in a few-shot setting. Recent advancements in large languages. The PGT framework tries to overcome this problem. The framework uses synthetic data generation, fine tuning smaller RAG models, and reinforcement learning. The proposed approach includes several key components: a synthetic generation pipeline to create training data, fine-tuning of a smaller RAG model comprising a dense retriever and a smaller LLM, training a Reward model to score domain-grounded answers, and aligning the RAG model using reinforcement learning. The authors here explore different innovative techniques such as modified retrieval procedures, Reinforcement Learning with Synthetic Feedback (RSLF), etc. The paper does not disclose results but it aims to compete with GPT-4 while using smaller models(10 billion parameters) for cost efficiency. The authors suggest that the cost efficiency can be 99.97% compared to GPT-3.5/4. The main contributions of this paper are the PGT framework, synthetic data generation, modified RAG structure, and RSLF approach.

This paper [19] of Saad-Falcon et al. introduced UDAPDR (Unsupervised Domain Adaptation via LLM Prompting and Distillation of Rerankers) which is an approach for adapting neural information retrieval models to new domains without the requirements of labeled data. The author has addressed the challenges of domain shifts in IR tasks where models that are trained on general datasets usually perform imperfectly on domains that are specialized. UDAPDR leverages large language models (LLMs) like GPT-3 and Flan-T5 XXL to generate synthetic queries for target domain passages. The method takes on a two-stage query generation process: first using an expensive LLM (in this case GPT-3) to create a small set of high-quality queries, later the queries are used to prompt less expensive LLM (Flan-T5 XXL) to generate larger numbers of queries efficiently. These synthetic queries are used to fine-tune multiple passage rerankers using DeBERTAV3-Large, which are then distilled into a single efficient ColBERTv2 retriever. The authors evaluated UDAPDR on various datasets including LoTTE, BEIR Natural Questions, and SQuAD while demonstrating significant improvements in zeroshot retrieval accuracy across various domains. For example, on the LoTTE pooled dataset, UDAPDR has improved Success@5 from 63.7% (zero-shot ColBERTv2) to 72.2% (using 10 rerankers with 10000 queries each). On the BEIR datasets, UDAPDR has achieved an average improvement of 5.2 points in nDCG@10 compared to zero-shot ColBERTv2. The highest reported accuracy was obtained on the BEIR Covid dataset, UDAPDR has achieved an nDCG@10 of 88.0. The authors had conducted tests to understand

how different factors like the number of rerankers, synthetic query counts and distillation methods affected the system's true performance. Notably, UDAPDR keeps query response times comparable to other retrieval systems but at the same time it focuses on the significant improvement of accuracy which makes it a strong choice for real-world use scenarios.

This paper [23] by Eschbach-Dymanus et al. explores the effectiveness of Large Language Models (LLMs) for domain-specific machine translation while focusing on business IT texts. In this paper, the authors have conducted a very comprehensive study to determine whether large language models (LLMs) could ever outperform their current neural machine translation systems. After testing a number of open-source models in both zero-shot and few-shot scenarios, including Llama-2, BLOOM, and Falcon, they concluded that Llama-2 13B was the most promising model for fine-tuning. In contrast to SAP's internal MT models, the authors have examined a variety of adaptation strategies in this work, ranging from prompting to complete fine-tuning. For testing and fine-tuning, the researchers used carefully selected SAP-internal parallel data, which included software user interface strings, training materials, user support messages, and commercial content. The researchers have used BLEU and COMET metrics to evaluate the performance on domain specific and general domain (FLORES) test sets. Many configurations of full fine-tuning were used here like LoRA (Low-Rank Adaptation), and QLora (Quantized Low-Rank Adaptation). These fine-tuning configs helped the researchers in analyzing effects of different configs on model performance and computational efficiency. The authors have also found out that full fine-tuning always outperformed parameter-efficient methods. When it comes to capturing domain-specific terminologies the outperformance is more noticeable. They have investigated the impact of training data volume and epochs on the model's performance. Insights into strategies focused on budget efficiency is provided in this study. The researchers have observed that when it comes to the general domain, LLMs could match SAP's MT models. On the other hand, even with extensive training it struggled when it came to closing the gap on domain-specific content. The highest reported accuracy score was achieved by the fully fine-tuned Llama-2 13B model which is trained on 400000 parallel segments. The score was 0.888 COMET / 0.686 BLEU on the SAP domain test set for English to French translation. The score of the LLM model advanced towards the performance of SAP's MT system on domain specific tasks. The research also included experiments on English to Japanese and English to Slovak language pairs. This comprehensive study has contributed many valuable insights into the challenges and the potential of adapting LLMs for translation tasks.

This technical report [33] by Siriwardhana et al. presents a very comprehensive study on domain adaptation of the Meta-Llama-3-70B-Instruct model for Securities and Exchange Commission (SEC) data. The researchers conducted extensive experiments to increase this model's domain-specific capabilities while mitigating catastrophic forgetting at the same time. Continual pre-training (CPT) with 70 billion tokens of SEC data had been employed in this case even though the report focused on an intermediate checkpoint that is trained on 20 billion tokens. The researcher team used Megatron-Core for distributed training on the AWS SageMaker HyperPod cluster, consisting of 4 nodes with 32 H100 (Nvidia Tensor Core GPU)

GPUs each. To mark the catastrophic forgetting, the researchers applied TIES merging techniques using MergeKit, combining the domain-adapted model with the original instruct model. Model performance on both domain-specific and general benchmarks was evaluated in the study. For domain-specific tasks, the researchers used subsets of TAT-QA and ConvFinQA, as well as financial classification and summarization tasks. General evaluations involved metrics from AGIEval, BigBench, GPT4all, and TruthfulQA. The researchers observed how significant improvements in domain-specific tasks can be noticeable after CPT and merging. For example, in financial classification, accuracy improved from around 60% (Llama-70B-Instruct) to nearly 100% after CPT, settling at about 90% after merging. In general, evaluations, while CPT initially led to some performance drops, merging helped recover much of the lost general capabilities. This study also measured perplexity on both domain-specific and general datasets. The researchers found that CPT reduced perplexity across all datasets. The highest domain-specific performance was achieved by the Llama-70B-CPT model before merging, while the Llama-70B-CPT-Merge model provided the best balance between domain-specific and general capabilities. For instance, when it came to the ConvFinQA task, the Llama-70B-CPT-Merge model achieved the highest score of approximately 0.62, which outperformed both the original Instruct model and the CPT model. The researchers also contributed by demonstrating the effectiveness of large-scale CPT on domain-specific data, comprehensive evaluation of domain-specific performance, and the successful mitigation of catastrophic forgetting through model merging. They also highlighted future directions for further improvements, such as exploring advanced alignment techniques and optimizing data processing methods for large-scale models.

This paper [8] by Farahani et al. provides a comprehensive review of domain adaptation, a subfield of machine learning that aims to address the challenge of distributional shifts between training and test data. The authors, Farahani et al., begin by explaining the motivation behind domain adaptation: in real-world applications, the assumption that training and test data come from the same distribution often doesn't hold, leading to performance degradation when models are applied to new domains. They thoroughly discuss the categorization of domain adaptation from various perspectives, including closed set, open set, partial, and universal domain adaptation. The paper then delves into different approaches to tackle domain adaptation, broadly categorizing them into shallow and deep methods. For shallow methods, they describe instance-based adaptation techniques like Kernel Mean Matching (KMM) and Kullback-Leibler Importance Estimation Procedure (KLIEP), as well as feature-based methods such as subspace alignment, transformation-based approaches like Transfer Component Analysis (TCA), and reconstruction-based methods like Robust visual Domain Adaptation with Low-rank Reconstruction (RDALR). The researchers have explored discrepancy-based methods in the deep domain adaptation section. Methods like Deep Adaptation Network (DAN), reconstruction-based approaches using autoencoders, and adversarial-based techniques were explored.The authors have provided detailed explanations of how these methods work and their underlying principles with mentioning each's strengths and limitations. Even though this paper does not present any new experimental results or accuracies but it does offer findings from numerous other studies in this field. This review serves as an excellent resource for researchers and practitioners looking to understand the land-

scape of domain adaptation techniques.

This paper by [6] Gururangan et al. in their study discussed adapting pre-trained language models for specific domains and tasks. The authors investigate whether it's beneficial to tailor a pretrained model like RoBERTa to the domain of a target task, even though such models are trained on diverse corpora. They conduct experiments across four domains (biomedical and computer science publications, news, and reviews) and eight classification tasks. The study introduces two main adaptation techniques: domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT). DAPT involves continued pretraining on domain-specific unlabeled data, while TAPT uses the unlabeled data from the target task. With a combination of these techniques, the authors proposed a method that performs selecting task-relevant data when additional unlabeled data in unavailable. For this study, RoBERTa was used as a base model and evaluated performance on tasks like CHEMPROT, RCT, ACL-ARC, SCIERC, HYPERPARTISAN, AGNEWS, HELPFULNESS, and IMDB. Domains that were the most distant fromRoBERTa's original training data surprisingly showed consistent improvements with DAPT. On the other hand, TAPT also showed its effectiveness in matching or exceeding DAPT's performance even after using less data. However, when these both are combined it achieved a macro-F1 score of 75.6. The score is quite improved compared to 63.0 for the base RoBERTa model. Data selection techniques like kNN-TAPT were also mentioned in this study to improve the performance of DAPT with less computational cost. The highest from this study was 95.6% on the IMDB dataset using the combination of DAPT+TAPT. The effectiveness of adaptive pretraining techniques was the main highlight of this study. Also, it provides practical strategies for improving model effectiveness on domain-specific tasks.

This paper [35] by Wang et al. a Context Enhancement and Domain Adaptation method for Textbook Question Answering (CEDA-TQA). This proposed method is empowered by Large Language Models (LLMs) and Retrieval Augmented Generation (RAG). Moving on, TQA could be a very complex challenge as it requires enough processing and understanding of different types of information such as text, diagrams, and formulas. However, with the proposed framework by the authors, they could manage to improve accuracy in education contexts of question answering. The power of LLMs and RAG took them to this point. A dataset named TQA was used in this research. It is an open-source science curriculum, specifically the earth science portion to manage resources related to computation focused. In their methodology, they preprocessed the dataset, formulated the TQA problem, and developed a framework for handling diagram and non-diagram questions. For the diagram related question authors especially employed a Multimodal Large Language Model. With the help of the mLLM, they extracted context from images and transformed them into non-diagram questions. Coming to the experiments, in this study the authors implemented the Mistral-7B-Instruct-v0.2 model, LoRA (Low-Rank Adaptation) method, and fine tunined using the PEFT (Parameter-Efficient Fine-Tuning) technique. Also, RAG was implemented for this study using the LanceDB vector database and benefited from efficient storage and retrieval of embeddings. With all these rigorous experiments the highest accuracy they got was 81.91% while both fine-tuning and RAG techniques were used together with the

Mistral-7B model. This research highlights the capabilities of medium-sized LLMs when it comes to specialized domains like education.

Setty, Sparty, and others, in their paper [29] aim to address the limitations of existing Retrieval Augmented Generation (RAG) pipelines for large language samples by answering questions about domain-specific documents, and thus ignoring financial documents and therefore have suboptimal incomplete references generated answers 9 or incorrect They can be incomplete. The study did not propose alternative sampling protocols; Instead, it sought ways to improve the reclamation of the existing RAG pipeline segment. These techniques include sophisticated chunking techniques that compute document features such as the table of contents, query expansion by inserting additional documents, modifying metadata descriptions for additional context, and algorithms that prioritize relevance over similarity when returning the Added testing, and domain-specific knowledge . While the paper did not specify the results because of the rapidly developing AI landscape, it noted that the proposed methods overcame some of the major limitations of the traditional RAG pipeline. Researchers used a variety of analytical parameters, such as page-level and paragraph-level accuracy for lists, contextual relevance, response fidelity, and LLM-based analyses for unstructured, aggregate objectives and to provide a comprehensive assessment of system performance, with baseline and shallow true data. Although the methods examined addressed some limitations, the paper acknowledged the potential for further improvement; The main challenge mentioned was recovering contextual data spanning multiple document types, which requires the development of a model to identify and integrate all relevant aspects to generate the response The authors suggest subscribed data-based knowledge accounts are not used, with fine tuning of input systems Potentially enhance the ability of the retrieval system to understand domain-specific nuances and retrieve relevant information in different aspects in many cases.

Nelson F. Liu et al. in their paper [27] talked about the performance of large language models (LLMs) while working with long input contexts. Specifically focusing on how these models handle information placed at different positions in lengthy texts. The motivation behind this study comes from the increasing importance of LLMs in applications that require processing long sequences. For instance, multi-document question answering and key-value retrieval. The authors in their research tried to understand whether LLMs can robustly use longer context. The reason behind this is that models like GPT-3.5-Turbo, Claude, and others that are trained on limited context windows (between 512-2048 tokens) are now deployed for tasks requiring much longer inputs. Also, they analyzed how different LLMs perform when the relevant information is located at the beginning, middle, or end of the input sequences through multiple controlled experiments. In addition, the authors wanted to test the models' ability to retrieve and use information from long contexts so they conducted experiments on multi-document question answering and key-value retrieval tasks. A key finding was the "U-shaped performance curve". In this curve, the models exhibited a primacy bias where models exhibited a primacy bias (best performance with information at the beginning) and recency bias (good performance at the end), while accuracy dropped significantly for information in the middle of long contexts. For example, GPT-3.5-Turbo got a substantial 20%

performance drop when accessing mid-context information. The study contributed by providing a comprehensive analysis of how LLMs use long contexts and proposing new evaluation protocols for future models. Coming to the performance part, GPT-3.5-Turbo achieved an accuracy of 56.1% in closed-book settings and 88.3% in oracle settings. The research concludes by saying that further improvements in how

Chapter 3

Dataset

3.1 Dataset Source

Our research leverages three distinct datasets shown in Figure:[3.1], each contributing unique perspectives to the landscape of legal information in Bangladesh:

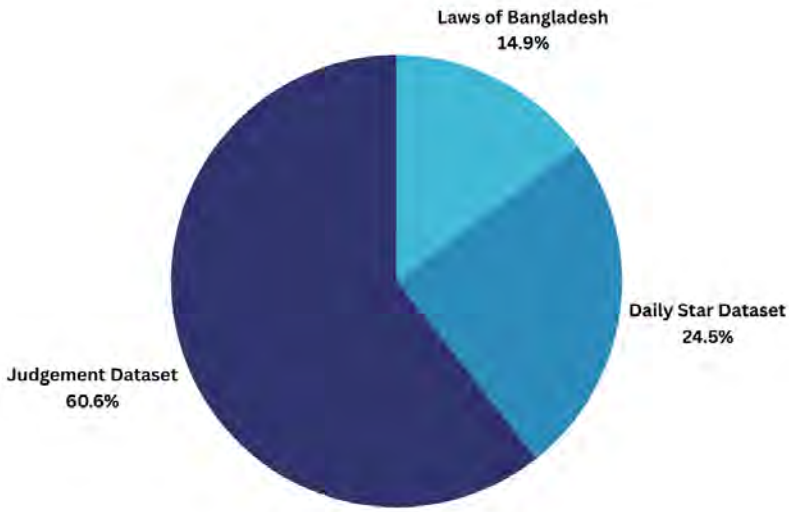


Figure 3.1: Distribution of Datasets

3.1.1 Laws of Bangladesh Dataset (1,380 entries)

At first, a model needs to know about the basic laws and regulations to precisely answer questions on those topics. So we developed a dataset called the 'Golden Dataset'. This dataset contains the official legal documents and rules which form the backbone of our legal system. We have scraped these laws from our official government website 'bdlaws.minlaw.gov.bd.'. This dataset contains constitutional laws, criminal statutes, civil regulations, etc shown in Figure:[3.2]. There are 1380 entries in this dataset which serves as the primary reference for legal interpretations. Also, we have generated a world could for this dataset shown in Figure:[3.3] that shows the most prominent words like "Act", "section", "Government", and "person".

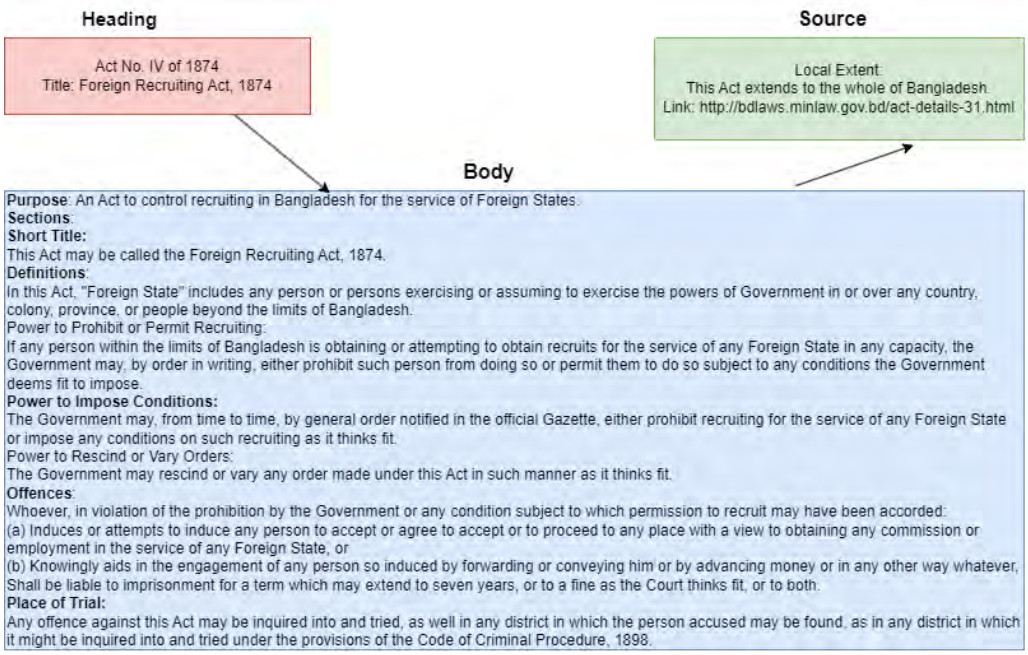


Figure 3.2: Laws of Bangladesh

The focus here is on legislative language, governmental authority, and individual rights or responsibilities in the country's laws.

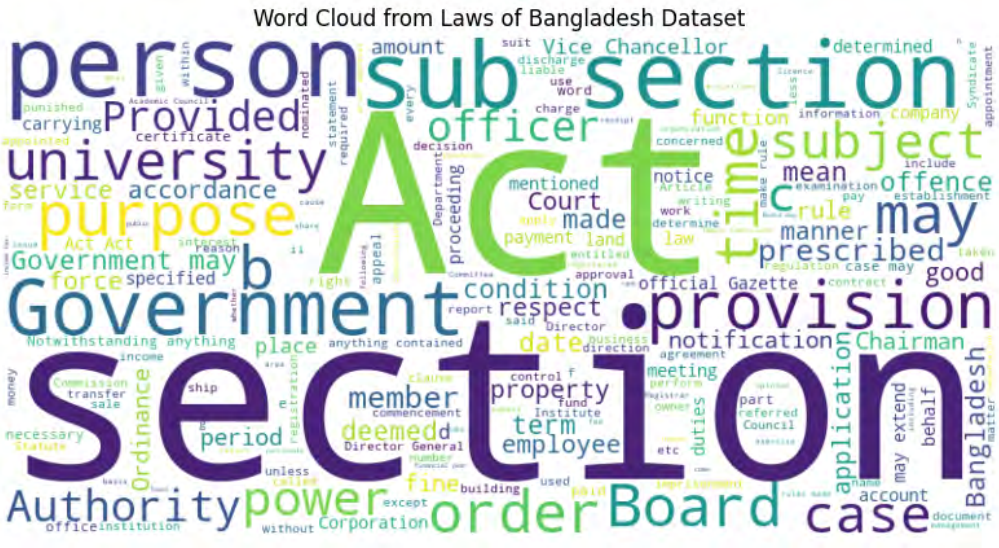
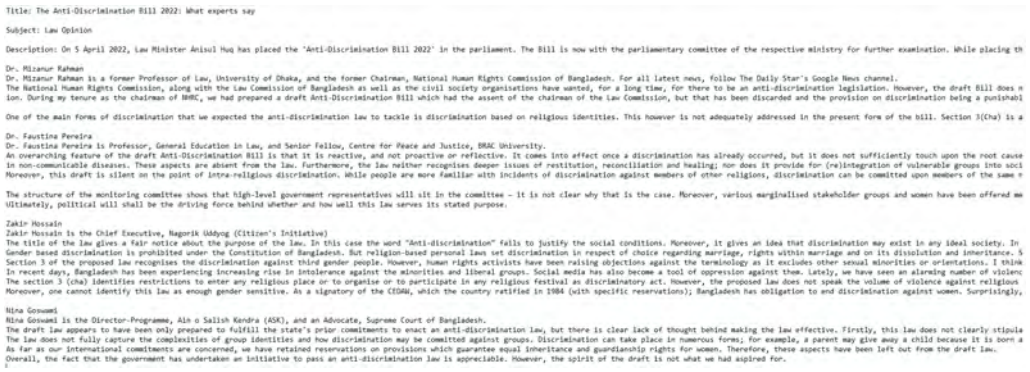


Figure 3.3: Word cloud for Laws of Bangladesh dataset.

3.1.2 Daily Star Dataset (2,276 entries)

So the first dataset is mostly single point based questions and answers. It has no context, no scenarios. It can not answer questions based on some specific context. So to train our model in some context based questions we created another dataset manually named The Daily Star dataset. We sourced our data from one of the

prominent newspapers of Bangladesh The Daily Star. It has a section where people would give questions based on their own context or scenarios and a prominent lawyer or a professor of law would give the answers based on his/her situation. This dataset is impressive for capturing the public discourse surrounding legal issues and providing context to the application of laws in real-world scenarios. We have 2276 entries here one of the examples is shown in Figure:[3.4].


Title: The Anti-Discrimination Bill 2022: What experts say

Subject: Law Opinion

Description: On 5 April 2022, Law Minister Anisul Huq has placed the 'Anti-Discrimination Bill 2022' in the parliament. The Bill is now with the parliamentary committee of the respective ministry for further examination. While placing th

Dr. Mizanur Rahman
Dr. Mizanur Rahman is a former Professor of Law, University of Dhaka, and the former Chairman, National Human Rights Commission of Bangladesh. For all latest news, follow The Daily Star's Google News channel.
The National Human Rights Commission, along with the Law Commission of Bangladesh as well as the civil society organisations have wanted, for a long time, for there to be an anti-discrimination legislation. However, the draft Bill does n
ion. During my tenure as the chairman of NHRC, we had prepared a draft Anti-Discrimination Bill which had the assent of the chairman of the Law Commission, but that has been discarded and the provision on discrimination being a punishabl

One of the main forms of discrimination that we expected the anti-discrimination law to tackle is discrimination based on religious identities. This however is not adequately addressed in the present form of the bill. Section 3(Cha) is a

Dr. Faustina Pereira
Dr. Faustina Pereira is Professor, General Education in Law, and Senior Fellow, Centre for Peace and Justice, BRAC University.
An overarching feature of the draft Anti-Discrimination Bill is that it is reactive, and not proactive or reflective. It comes into effect once a discrimination has already occurred, but it does not sufficiently touch upon the root cause
in non-communicable diseases. These aspects are absent from the law. Furthermore, the law neither recognises deeper issues of restitution, reconciliation and healing; nor does it provide for (re)integration of vulnerable groups into soci
Moreover, this draft is silent on the point of intra-religious discrimination. While people are more familiar with incidents of discrimination against members of other religions, discrimination can be committed upon members of the same r

The structure of the monitoring committee shows that high-level government representatives will sit in the committee – it is not clear why that is the case. Moreover, various marginalised stakeholder groups and women have been offered ma
Ultimately, political will shall be the driving force behind whether and how well this law serves its stated purpose.

Zakir Hossain
Zakir Hossain is the Chief Executive, Nagorik Uddyog (Citizen's Initiative)
The title of the law gives a fair notice about the purpose of the law. In this case the word "Anti-discrimination" fails to justify the social conditions. Moreover, it gives an idea that discrimination may exist in any ideal society. In
Gender based discrimination is prohibited under the Constitution of Bangladesh. But religion-based personal laws set discrimination in respect of choice regarding marriage, rights within marriage and on its dissolution and inheritance. S
Section 3 of the proposed law recognises the discrimination against third gender people. However, human rights activists have been raising objections against the terminology as it excludes other sexual minorities or orientations. I think
In recent days, Bangladesh has been experiencing increasing rise in intolerance against the minorities and liberal groups. Social media has also become a tool of oppression against them. Lately, we have seen an alarming number of violenc
The section 3 (cha) identifies restrictions to enter any religious place or to organise or to participate in any religious festival as discriminatory act. However, the proposed law does not speak the volume of violence against religious
Moreover, one cannot identify this law as enough gender sensitive. As a signatory of the CEDAW, which the country ratified in 1984 (with specific reservations); Bangladesh has obligation to end discrimination against women. Surprisingly,

Nina Goswami
Nina Goswami is the Director-Programme, Ain o Salish Kendra (ASK), and an Advocate, Supreme Court of Bangladesh.
The draft law appears to have been only prepared to fulfill the state's prior commitments to enact an anti-discrimination law, but there is clear lack of thought behind making the law effective. Firstly, this law does not clearly stipula
The law does not fully capture the complexities of group identities and how discrimination may be committed against groups. Discrimination can take place in numerous forms; for example, a parent may give away a child because it is born a
As far as our international commitments are concerned, we have retained reservations on provisions which guarantee equal inheritance and guardianship rights for women. Therefore, these aspects have been left out from the draft law.
Overall, the fact that the government has undertaken an initiative to pass an anti-discrimination law is appreciable. However, the spirit of the draft is not what we had aspired for.

Figure 3.4: Laws of Bangladesh

Also, we have created a word cloud shown in Figure:[3.5] for this dataset Prominent words like "Bangladesh", "case", "law", "right", and "court" suggest a focus on legal and political news coverage, with emphasis on national issues and governance.

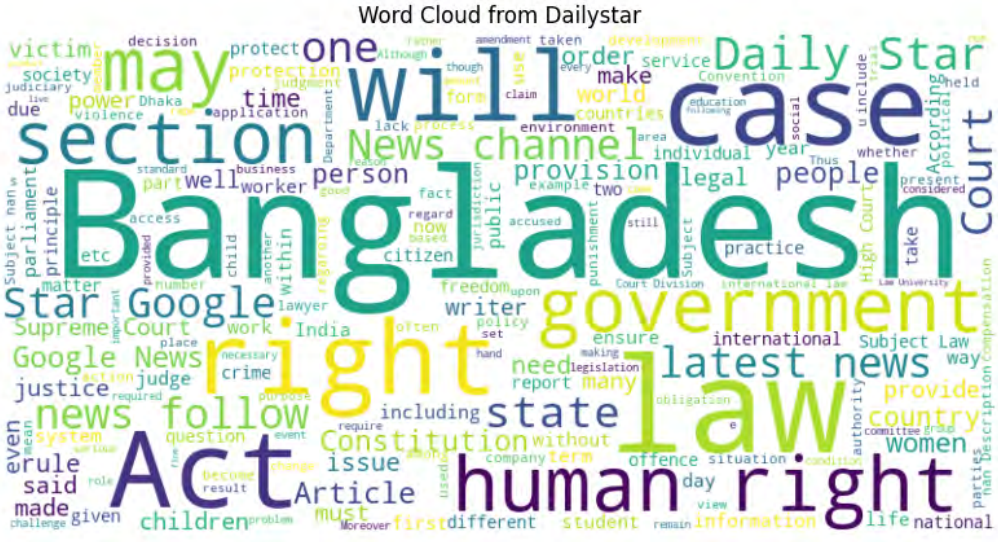


Figure 3.5: Wrod cloud for Daily Star dataset.

We got 455 categories of subjects from the Daily Star newspaper from which we took 9 prominent subjects to showcase the data, shown in Table:[3.1].

3.1.3 Judgement Dataset (5,633 entries)

Our model still was not good enough to answer all the context based questions. Sometimes it hallucinated and gave irreversible answers. So to solve this problem we created another dataset manually named the Judgement dataset. There are a lot of cases and their judgments online but most of them are copyrighted. So we looked

for open source cases and manually created a dataset with 5633 entries of text files. These judgments are a testament to the Bangladeshi judiciary. This dataset provides insights into the interpretation and application of laws. The dataset includes case name, case category, case note, and judgment shown in Figure:[3.6]. This also helps the model to give relevant information and relevant cases that have the same context as the users. We have also generated a word cloud shown in Table:[3.7] for this that shows prominent legal terms such as "case," "petitioner," "respondent," and "section" indicating frequent references to legal proceedings and court rulings.

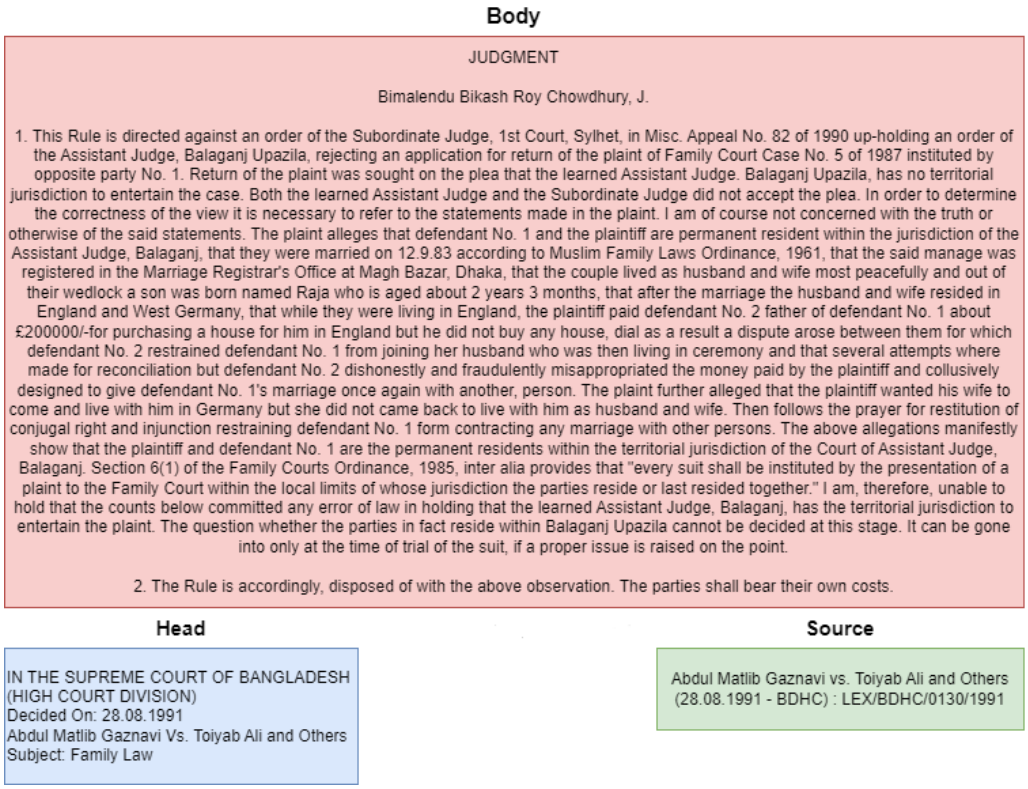


Figure 3.6: Judgement Data

This multifaceted analysis not only validates the richness and diversity of our datasets but also provides crucial insights that inform the development of our AI model, ensuring its relevance and effectiveness in the Bangladeshi legal context.

3.2 Data Collection Tools

Our research includes web scraping techniques to gather information efficiently and systematically:

- **Selenium:** Selenium was instrumental in automating web browsers, allowing us to navigate complex web pages and extract data from the Internet. Selenium's ability to interact with web elements was particularly useful for accessing legal databases and news archives like the Daily Star.
- **BeautifulSoup4 (BS4):** We utilized BS4 for HTML parsing, extracting relevant information from the scraped web pages.

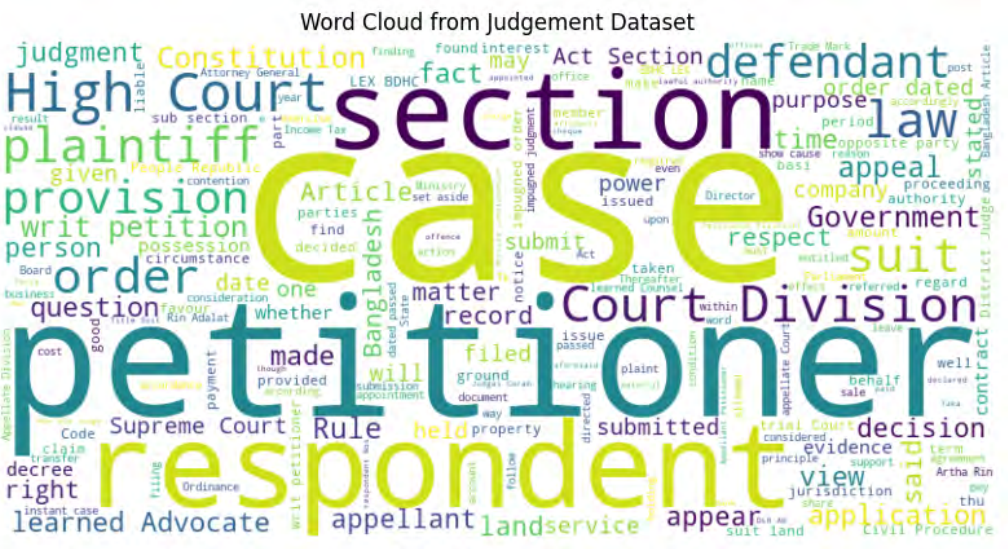


Figure 3.7: Word cloud judgment Data

- **Pandas Library:** The data we got from BS4 parsing was then saved to Pandas DataFrame. Then we have extracted the data from DataFrame and saved it in a .csv file. Then again using Pandas Library we converted those data in .txt files.

3.3 Data Translation

Bangla Neural Machine Translation (BanglaNMT) is an NMT system developed by BUET CSE. It was trained on various Bangla datasets. It is designed to translate Bangla to English and English to Bangla. It is specialized in preprocessing techniques that really work well with complex documents. In our three datasets, many of the data were in Bangla. So we have used this translation model to convert those data to English for better understanding and also deleted Bangla stop words. Before translation, the data looked like Figure:[3.8] and after translation the data looked like Fiture:[3.9].

3.4 Data Preprocessing

Our goal is to maintain the best quality and most usable data possible. It is very important for RAG systems. Often data contains a lot of noise such as HTML tags, code snippets, and special characters. This can cause difficulties for the retrieval system to generate accurate data. Our preprocessing method addresses these problems while maintaining the original form of the documents.

3.4.1 Preprocessing Method

There are three types of methods we used. Those are (I).HTML tag removal, (II). Code snippet removal, (III). Special character cleaning. To further explain,

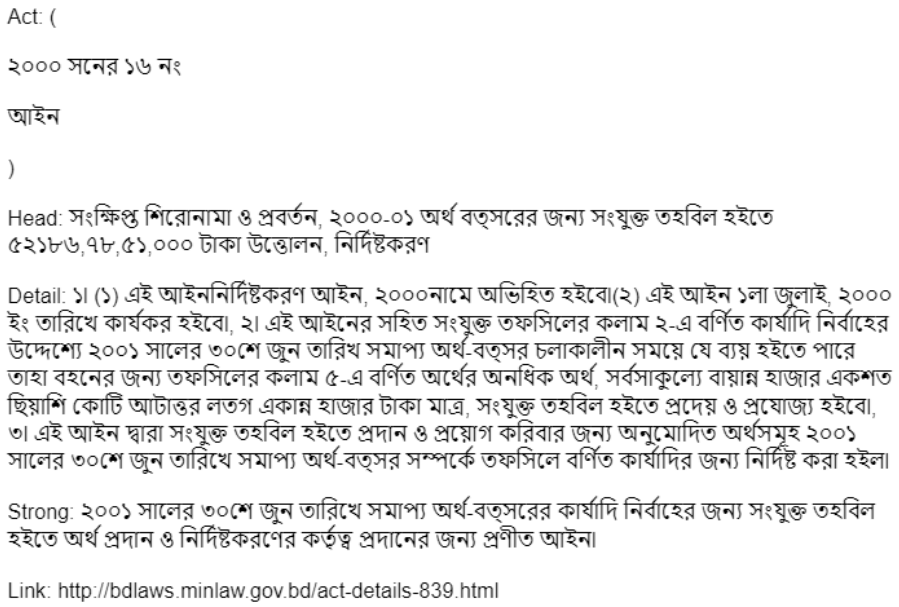


Figure 3.8: Before Translation

I. HTML Tag Removal

The BeautifulSoup (BS4) is used to remove any HTML content from the datasets. It removes the tags but maintains the original text structure. It specifically targets the tag
 and <p>. These two represent line breaks and paragraphs. This BS4 code snippet replaces these tags with a new line (\n). By doing this it maintains the original text formatting. After removing it returns the texts without any HTML content. It also ensures that the text contents are structured and readable.

II. Code Snippet Removal

To remove code snippets from any documents we used regular expressions (regex). It generally removes two types of codes - i. multi-line code blocks and inline code blocks. The first regex which is (r'“‘[\s\S]*¿“') eliminates any data surrounded by triple backticks(‘ ‘ ‘). And the second regex (r'[\n]+') matches and nullifies inline code surrounded by single backticks(‘). After the removal process of both types of code blocks, it returns the text without any code snippets.

III. Special character cleaning

Using a regular expression we also remove special characters from text documents but we keep some of the punctuation that is necessary. We split the text documents into single lines and using a regular expression we remove any special characters. Also, it removes any extra space between words if there is any. After removing the special character and extra spaces, the lines are joined together again.

Then we combine these three components into a single preprocessing function. And now the data is clean, readable, and free from any noise elements.

Act: (Act No. XVI of 2000)

Head: Short title and commencement, withdrawal of Tk 521867851,000 from the consolidated fund for the 2000-01 financial year, specification

1. Details:

(1) This Act may be called the Defining Act, 2000.

(2) This Act shall come into force on July 1, 2000.

(3). The money approved by this Act to be paid and applied from the attached fund is specified for the activities specified in the Schedule for the financial year to be expired on the 30th June, 2001.

Strong: An Act to give authority to pay and specify funds from the consolidated fund for the expiry of the financial year on 30 June 2001. Link:
http://bdlaws.minlaw.gov.bd/act-details-839.html

Figure 3.9: After Translation

3.5 Document Processing

We apply these preprocessing methods to every document. Then we save that clean content with the original file names. So first the method retrieves documents from the list and processes them using the above mentioned preprocessing function. It eliminates any kind of HTML tags, special characters, and code snippets. Then it retrieves the original document name from metadata creates an output file with that same name and saves the clean text.
Our preprocessing method effectively cleans text documents for RAG systems by removing HTML tags, code snippets, and special characters while preserving document structure and original file names. This approach ensures that the cleaned documents maintain their integrity and organization, making them suitable for effective retrieval and generation tasks in RAG systems

3.6 Semantic Chunking

In the semantic chunking process text is divided into smaller, meaningful segments on the basis of content rather than arbitrary measures like character count. With the help of this advanced method, embedding-based models it becomes easy for models like HuggingFaceBgeEmbeddings to generate vector representations of text while capturing semantic information. The most interesting part of this method is that whenever there is a significant shift in meaning the method identifies a breakpoint. This finding of a breakpoint is done by using statistical techniques like standard deviation. This helps to keep the semantically coherent ideas to remain together while different topics are separated. The benefits of this method are that it helps in improved context preservation and creating adaptive chunk sizes eventually allowing complex ideas to remain intact and distinct sections to break-down. However, computational-wise it is quite intensive as it requires processing the text through a language model to generate embedding. Nonetheless, for any task that needs contextual understanding like summarization or question-answering, with semantic chunking a better result can be achieved when compared with simple character

based methods. From a word cloud shown in Figure:[3.10], from that we can see that frequently occurring legal terms and phrases, such as ”section,” ”case,” ”order,” and ”Penal Code,” suggesting a focus on legal documents or court proceedings.

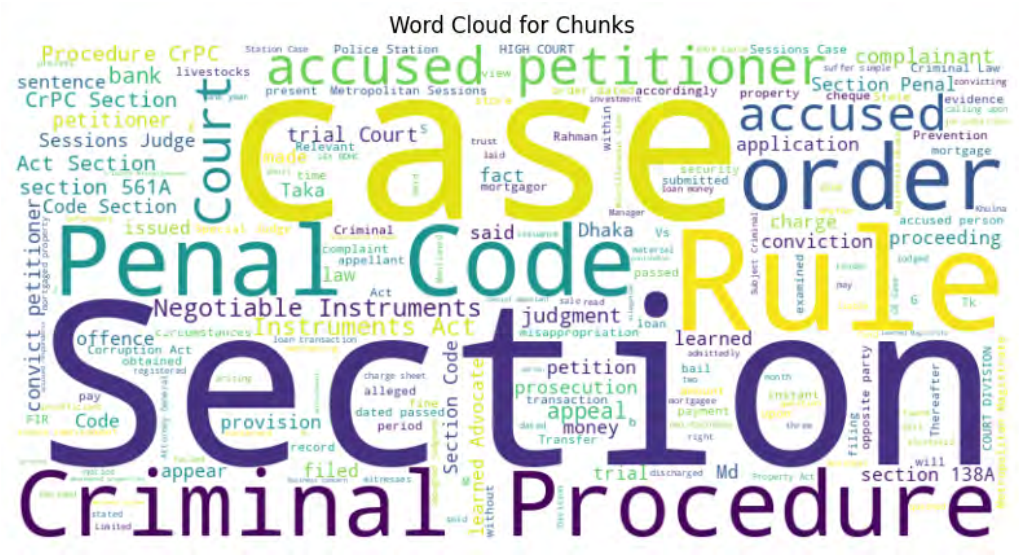


Figure 3.10: Word cloud after chunking.

Main Category	Subcategory	Count
1. Legal Analysis and Review	Law Review	5
	Law Analysis	22
	Law in Depth	4
	Judgment Review	7
	Law Opinion	89
	Constitutional Analysis	3
2. Human Rights and Advocacy	Rights Watch	63
	Rights Advocacy	44
	Human Rights Advocacy	5
	Rights Corner	11
	Writing for Equality	4
	Women's Rights	1
	Child Rights	1
	Climate Justice	1
	Technology and Human Rights	1
	Law Reform	10
	Law Campaign	5
3. Specialized Legal Areas	Law and Gender	1
	Environmental Law	1
	Intellectual Property Rights	1
	Business and Human Rights	1
4. Legal Education and Awareness	Law Interview	17
	Legal Education	3
	Legal Awareness	1
	For Law Students	3
5. Current Affairs and Updates	Year-End Judgment Review	1
	Year-End Law Review	1
	New Law	1
	Global Law Update	1
	Law News	37
6. Legal System and Institutions	Court Corridor	6
	Lower Courts	1
	Parliament Scan	4
7. Special Features	Constitution Day Special	2
	World Human Rights Day Special	3
8. Legal Commentary	Law Observation	1
	Law Letter	61
	Law Vision	47
	Reviewing the Views	18
9. Miscellaneous	Book Review	18
	Law Event	48
	Your Advocate	17
	People's Voice	7

Table 3.1: Categories of Legal Articles in The Daily Star Newspaper

Chapter 4

Model

4.1 Transformer

The transformer architecture shown in [4.1], Large Language Models and Natural Language Processing is considered a holy trinity in the field of AI research. It especially works in the domain of understanding context and text generation. [9] These transformer models have their own high-level self-attention mechanism. It enables them to capture long dependencies within text. [36]It is also very good at efficient learning and fast computation. Large Language Models like Mistral 7B work very well on the architecture of transformers. It utilizes the models by training a huge amount of extensive textual documents. Then The LLMs can learn comprehensive knowledge by self-supervised learning. Then we can fine tune the LLM for domain specific tasks like question answering and classification. The specification of Mistral 7B is shown in Figure:[4.2]

4.2 Mistral7B

Mistral 7B is one of the biggest achievements in the Large Language Models world. It has a massive 7.3 billion parameters. [15] It is specially fine tuned to deliver high performance in different Natural Language Processing (NLP) related tasks. Although its scale is massive, it can adapt to different tasks and fields. Mistral 7B can reach the levels of Codelama 7B on coding tasks while maintaining proficiency. Two of the core innovations of Mistral 7B are i. Grouped Query Attention(GQA) and ii. Sliding Window Attention(SWA) mechanisms. GQA helps to boost the speed of processing queries faster and more efficiently. On the other hand, SWA solves the problem of handling longer sequence texts. It allows the model to enter hidden states with a little computation cost.
Mistral 7B's architecture shown in Figure:[4.3]improves Mistral's capabilities as well as helps to process documents faster with speed, accuracy, and efficiency.
At the heart of it, Mistral 7B utilizes transformer architecture. Its self-attention mechanism makes the sequence processing very easy. The architecture is made of various layers of self attention and feed forward neural networks. Every layer's self attention mechanism allows the model to work with different types of context and information. And by this, it holds a contextual relationship between the long sequences. The model gives an attention score to each word in the sequence and assigns importance levels to those words based on their contextual relevance. The

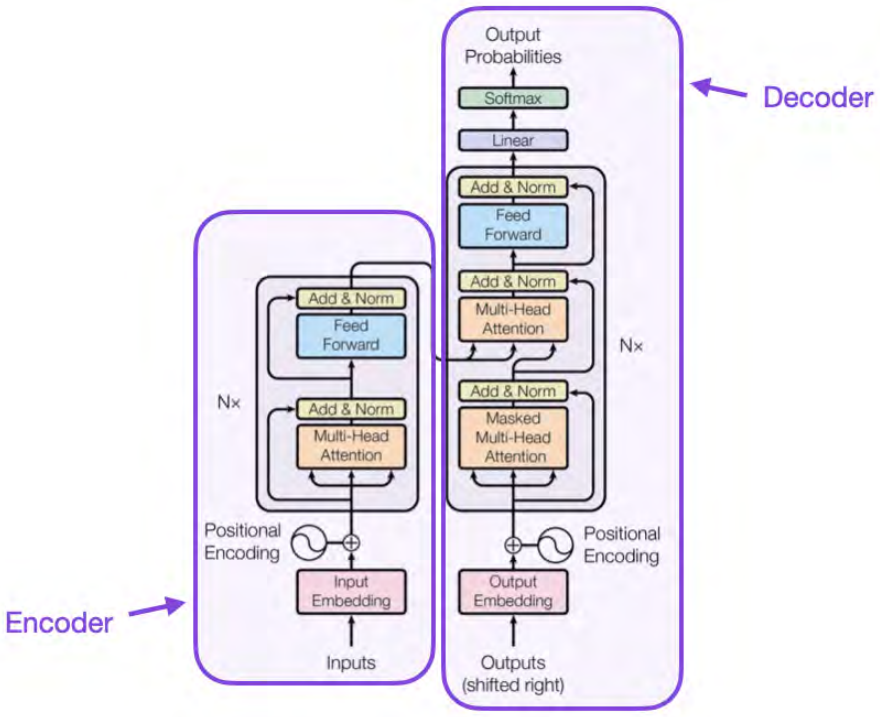


Figure 4.1: Transformer Model Architecture

model also used its feed forward neural network to process the information it gets from self-attention. Mistral 7B works really well in the fields of reasoning, knowledge, and reading comprehension. This makes it a fine choice for NLP tasks.

4.3 Llama 3.1

Llama 3.1 developed by Meta is another milestone in LLMs world. It is fine tuned with billions of parameters for specifically high performance and to balance speed, accuracy, and efficiency. It is specialized in both general knowledge and domain specific fields. It also has introduced attention mechanisms that make it very versatile for handling complex tasks like text generation, summarization, and translation. The main version has a massive 405B parameters and the other two versions have 80B and 8B parameters respectively. [22]. The architecture of the model is shown in Figure:[4.4]
It has several key specifications. It introduces highly improved attention layers that help it to focus on important parts of the input. This helps it to handle longer texts and makes a great relationship between words and concepts. Because it is fine tuned so greatly it always gives higher accuracy. The most impressive feature is its ability to adapt to certain domain specific fields like medical, legal, and technical fields. It is also based on transformer architecture which is now a standard for NLP models. The architecture's self-attention mechanism helps to get the context of different parts of the input sequence. Then the feed forward neural networks process the information they get from self-attention layers and enable efficient transformation. It excels in tasks such as reading comprehension, commonsense reasoning, text summarization, and code generation. It also has the ability to control more complex

Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
window_size	4096
context_len	8192
vocab_size	32000

Figure 4.2: Mistral 7B Specification

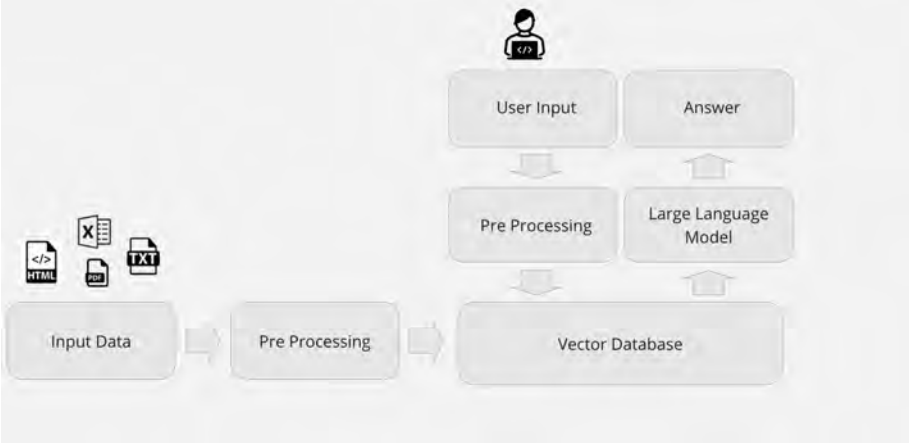


Figure 4.3: Mistral 7B Architecture

datasets that produce contextually relevant responses. The implementation of this model's process is shown in Figure:[4.3].

4.4 Deci AI

Deci AI is an advanced artificial intelligence platform that is designed to optimize deep learning models for improved performance and efficiency. Deci AI considerably improves speed and accuracy while maintaining the high quality of the model by tailoring it for various hardware and deployment settings through the use of automated neural architecture search (NAS) and optimization approaches. [25] The platform automates the process of designing and training deep learning models which enables the developers to achieve superior inference performance with reduced computational costs. Deci AI is a highly versatile platform that can be used across many industries, from computer vision to natural language processing (NLP). It helps create efficient models that handle complex tasks in real time. Deci AI has become an

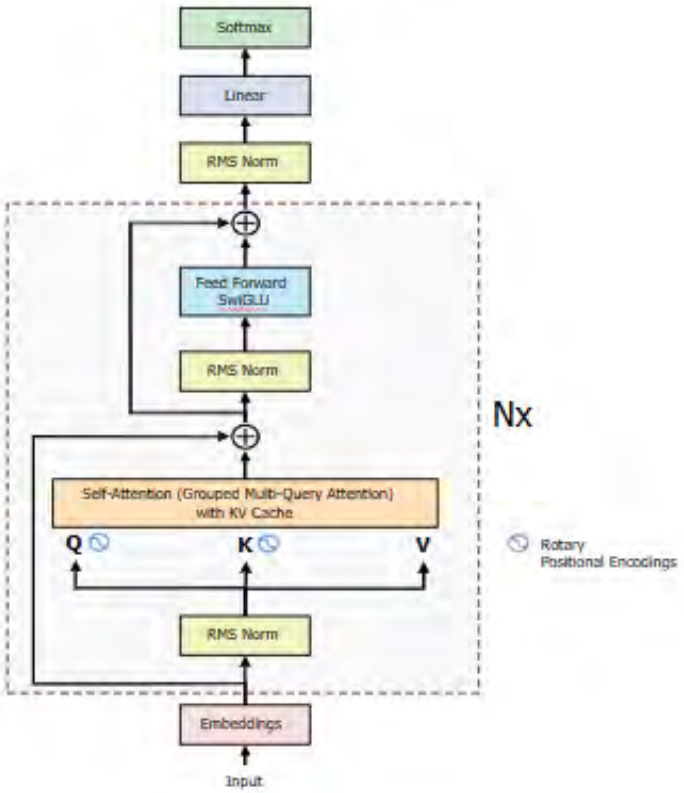


Figure 4.4: Llama 3.1 Architecture

essential tool for organizations that are looking to deploy AI at scale while using fewer resources.

4.5 SOLAR-10.7B-v1.0

SOLAR-10.7B-v1.0 is a cutting-edge large language model which is developed by Upstage. This LLM is a significant leap forward when it comes to natural language processing capabilities. This advanced model is built on the foundation of the transformer architecture. This model is both powerful and efficient at the same time. It consists of 10.7 billion parameters. [24]It has a very good balance between computational power and high performance. SOLAR-10.7B-v1.0 language model uses Stability-Oriented Language Refinement. It is a technique that is used to make the model better at almost all kinds of language tasks. Stability-Oriented Language Refinement keeps the language tasks stable and reliable at the same time. The Stability-Oriented Language Refinement methodology enhances adaptive learning, encompassing meticulous data curation, and advanced training paradigms mechanisms. The architecture of this model is shown Figure:[4.6].

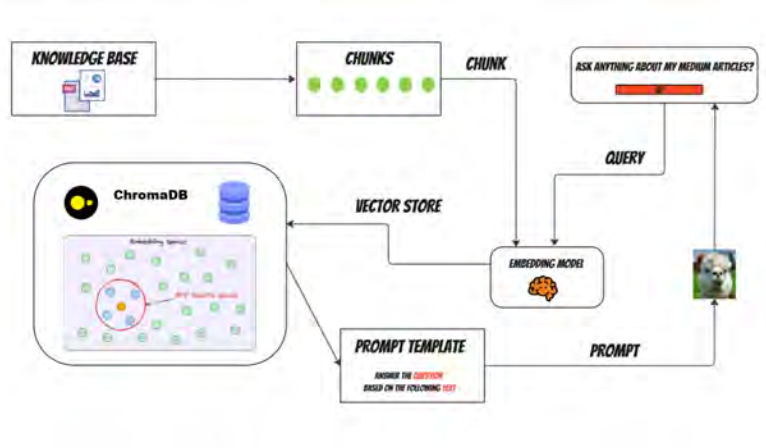


Figure 4.5: Mistral 7B Implementation

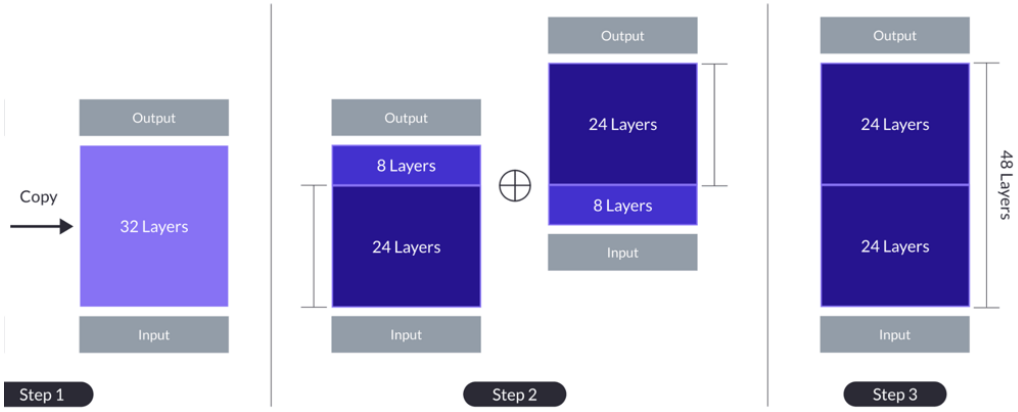


Figure 4.6: SOLAR-10.7b-v1.0 Arthitecture

4.6 Embedding: mxbai-embed-large-v1

The mxbai-embed-large-v1 is a groundbreaking creation from the innovative mixedbread-ai company that stands as a paragon of excellence in the domain of semantic representation. This cutting-edge embedding solution is precisely engineered to excel when it comes to capturing the nuanced semantics of textual data across different spectra of domains. The mxbai-embed-large-v1 model uses the latest neural network technologies to demonstrate the core essence of language in complex (high-dimensional vector spaces) ways to facilitate an unmatched accuracy score in tasks like comparing meanings (semantic similarity tasks) and searching information (information retrieval applications).[26] This model's architecture or design is based on advanced attention mechanisms that have new features to help pick up on subtle changes in context. Another feature is understanding a long range of connections in the texts with great accuracy. As the model is trained on a very large and carefully chosen dataset, it understands everyday language and specialized jargon which makes it a very highly adaptable model in many different fields. Its capacity to produce intricate, context-aware representations enhances the efficiency of a range of natural language processing tasks and provides opportunities for more sophisticated applications, such as text analysis, sentiment analysis, and content recommendation. Optimized for efficient processing, the model offers a strong balance between

computational speed and quality of its language representations which makes it valuable for both research and practical use cases that require a deep understanding of language on a large scale.

4.7 Bangla NMT

BanglaNMT is a state of the art Neural Machine Translation system which is developed by the NLP group at Bangladesh University of Engineering and Technology (BUET). The system addresses the historically low-resource status of Bangla language translation through a comprehensive parallel corpus of 2.75 million Bangla-English sentence pairs[7]. The system uses OpenNMT's implementation of the Transformer model with specialized preprocessing including a custom sentence segmenter and a novel "aligner ensembling" approach with batch filtering. The key innovations were: (1) a customized Bengali sentence segmenter, (2) aligner ensemble combining multiple aligners with LASER-based filtering, and (3) batch filtering for efficient noise removal [7]. The model uses the SentencePiece tokenizer with subword regularization and a 32k vocabulary for both languages. The reason we used BanglaNMT as our translation model is because their approach was particularly effective for legal documents due to their unique solution. As per authors, standard aligners struggle with legal texts because they contain bullet points ending in semicolons that would incorrectly fragment sentences. To address this, the authors developed a specialized regex-based segmenter and aligner specifically for legal documents. Also, they trained their model on 28,218 law related sentence pairs. These sentences were sourced from The Legislative and Parliamentary Affairs Division of Bangladesh website and "Heidelberg Bangladesh Law Translation Project"[7]. This represents about 1% of their total dataset of 2.75 million sentence pairs. As a result, their model achieved significant improvements, surpassing previous approaches by more than 9 BLEU points on the SUPara-benchmark test set (32.10 vs 22.68 for Bengali-to-English)[7].

4.8 Domain Adaptation

A base language model like Llama 3.1 8B, while powerful and broadly knowledgeable, is not ideal for specialized legal queries. Without fine-tuning on the legal corpus the model would suffer due to the lack of in-depth domain-specific knowledge, outdated legal information, misinterpretation of complex legal concepts, inability to replicate specific legal reasoning patterns, and inconsistency in responses to similar legal questions [9]. Additionally, general models may struggle with jurisdiction-specific information, have a higher risk of hallucinating in specialized contexts. Fine-tuning the model on a specific legal corpus addresses these limitations, making it more accurate and reliable. To address this issue we load base Large Language Models - either Llama-3.1-8B-Instruct RAG or Mistral 7B v0.3 RAG - using the Transformers library along with their corresponding tokenizers. To optimize for memory efficiency and performance, we employ the bitsandbytes library for quantization. We employed a Parameter efficient Fine-tuning technique called Qlora. Then we opted for an approach to unsupervised self-training on our laws of Bangladesh dataset without explicit labels or supervised fine-tuning targets [19]. The process begins by

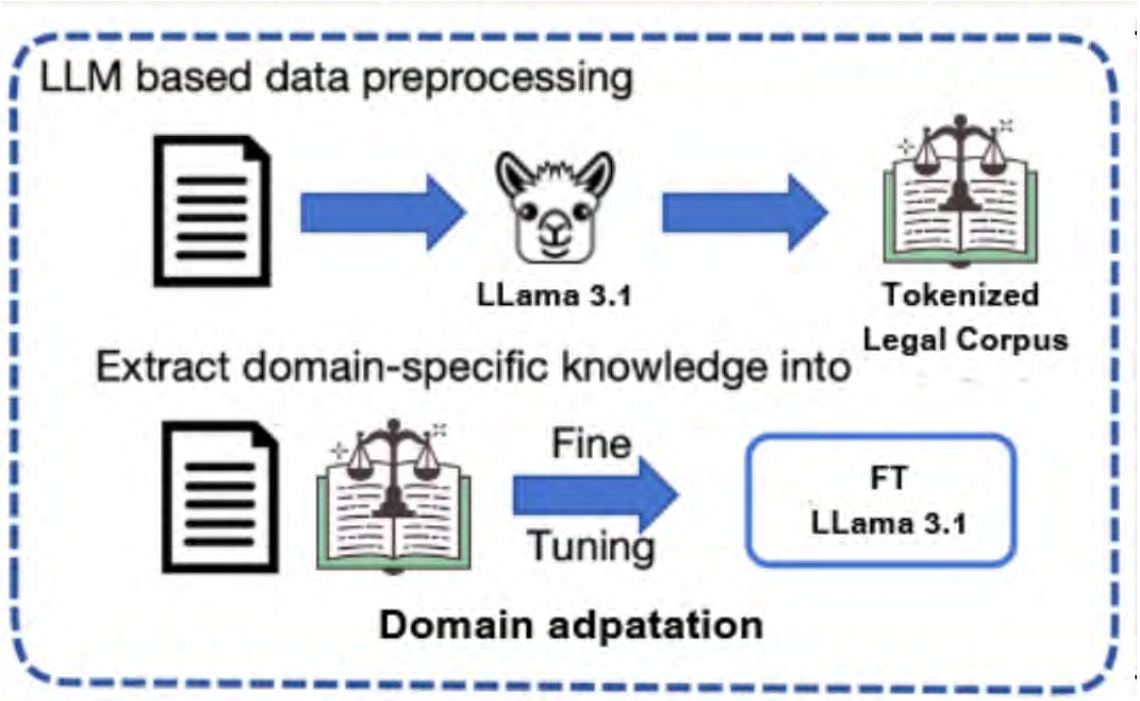


Figure 4.7: Domain adaptation process

preprocessing the legal texts, tokenizing them, and creating a dataset.Through multiple epochs of training on our unlabeled legal corpus, the model gradually improves its understanding and generation capabilities for legal content, essentially teaching itself through repeated exposure and prediction tasks[31]. The whole process is demonstrated in the Figure:**xxx**.

4.9 Full Analysis of BanglaLawBot

At first, our system loads all the appropriate libraries to cover all the various functionalities including text processing, document handling, vector storage, embedding generation, language modeling, and the overall pipeline for retrieval-augmented generation. The system combines components from Hugging Face's Transformers, LangChain, and PyTorch to create its legal information retrieval and response generation capabilities. At first, we combined all three of our datasets dailystar, judgment, and laws of Bangladesh which we named as golden dataset, and store them in one folder. Then we create a dictionary of keyword arguments for the text loader. While loading the document we make sure the text loader can detect the text files encoding and load them. Then we actually execute the loading process, reading all the ".txt" files in the specified directory and its subdirectories, and storing the resulting documents in the docs variable. Then using a specialized text splitting tool from the langchain_experimental.text_splitter module we split the docs variable into manageable chunks. However, Unlike simple text splitter that split text based on character or word count, SemanticChunker splits texts in a way that preserves the semantic meaning of our documents. This is done by mxbai-embed-large-v1 by mixedbread-ai which is used to understand the semantic content of the text. Using "standard_deviation" means the fragments would analyze the semantic similarity

of sentences and create breaks where there are significant deviations in meaning. This approach solves the challenge that the traditional chunking methods face. Traditional chunking processes are incapable of preserving context and coherent ideas within each segment. However, our method of chunking preserves contextual semantic insights which can be particularly useful for processing legal documents where maintaining the integrity of legal concepts is crucial.
Once the documents are chunked, each chunk needs to be converted into a vector representation. This is done using our mxbai-embed-large-v1 embedding model. The embedding process works as follows: a) The text chunk is fed into the embedding model. b) The model processes the text and outputs a high-dimensional vector (usually hundreds or thousands of dimensions). c) This vector captures the semantic meaning of the text in a way that's understandable to machines. After embedding, these vectors are stored in Chroma vector database for quick retrieval

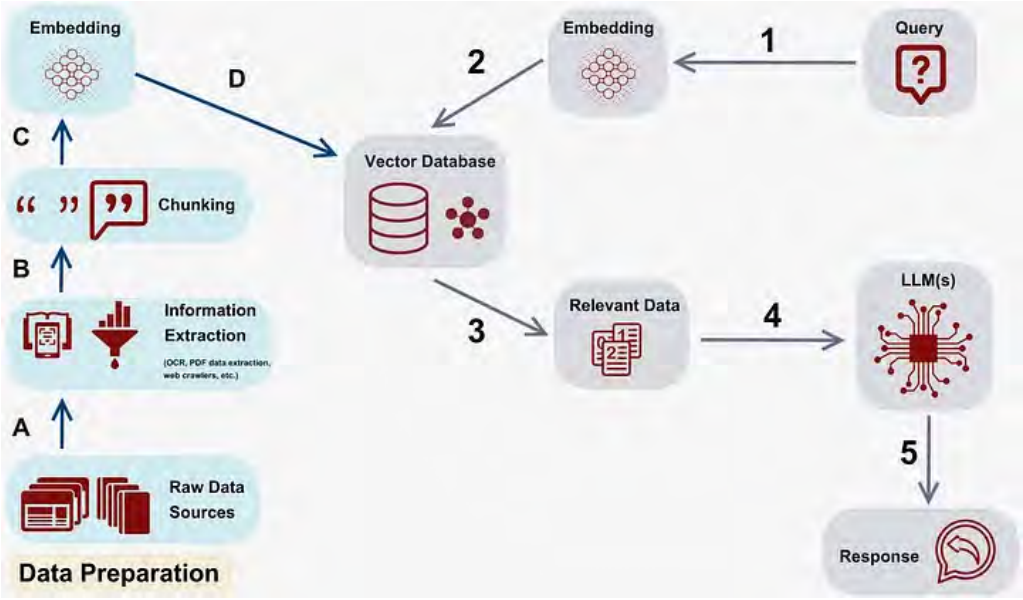


Figure 4.8: RAG Architecture.

After That, we load our proposed Large Language Models - either Llama-3.1-8B-Instruct RAG or Mistral 7B v0.3 RAG which were fine tuned using our legal dataset which is mentioned in the domain adaptation part of our paper. After loading the model we create a text generation pipeline, using the pipeline function from the Transformers library. This pipeline encapsulates the process of converting input text to tokens, passing those tokens through the model, and decoding the output back into text. Furthermore, we lower the temperature of our model to get a more deterministic output from the model.
When a query is made, the system can quickly search through these pre-computed embeddings to find the most relevant chunks, without needing to re-process the original documents. The system does that by calculating the similarity between the query embedding and all the stored document chunk embeddings using cosine similarity. The system then ranks the document chunks based on their similarity scores and retrieves the top k most similar chunks. Then by importing the PromptTemplate function using the langchain library we send all the corresponding context of the respected query and a set of instructions to guide the model's response, helping

it understand the context and generate relevant and coherent language-based output. Then this generated output is reviewed by an expert preferably by a lawyer for verification. Keeping a human in the loop in the heart of our model. Once verified, the output is sent to the user and also stored in a vector database to enhance future responses. This process ensures accuracy, relevance, and continuous improvement of the system's performance by leveraging expert knowledge and building a growing repository of verified information[14].

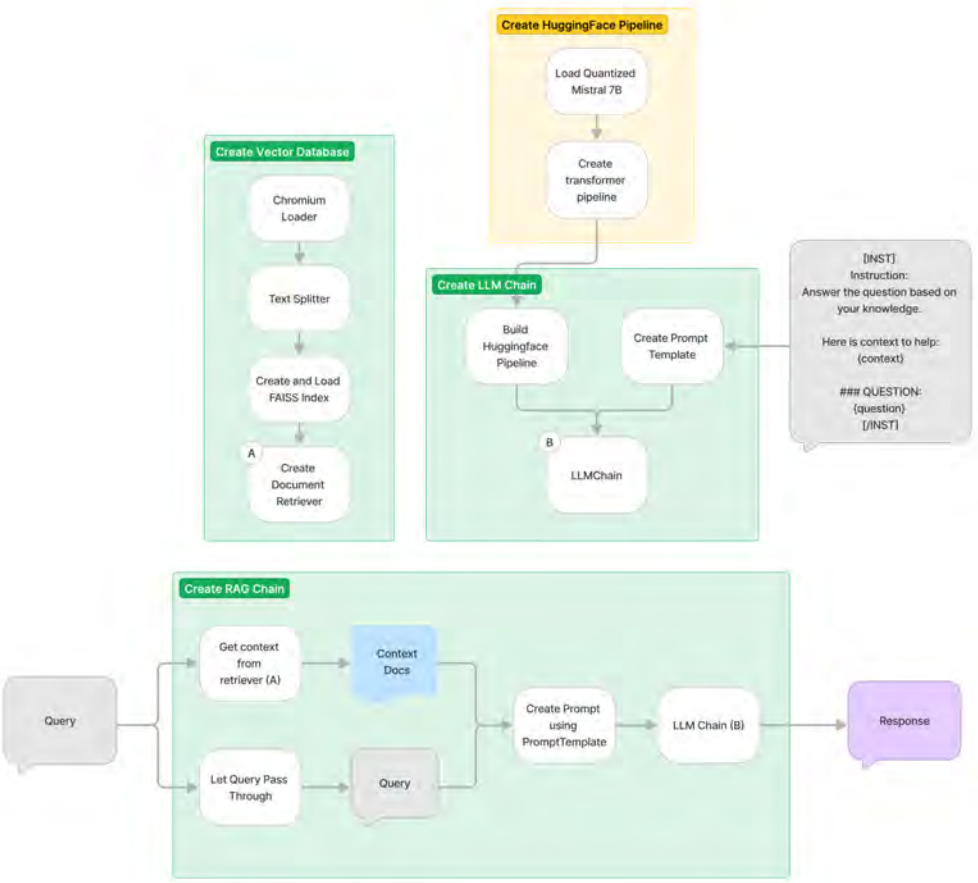


Figure 4.9: Model Pipeline.

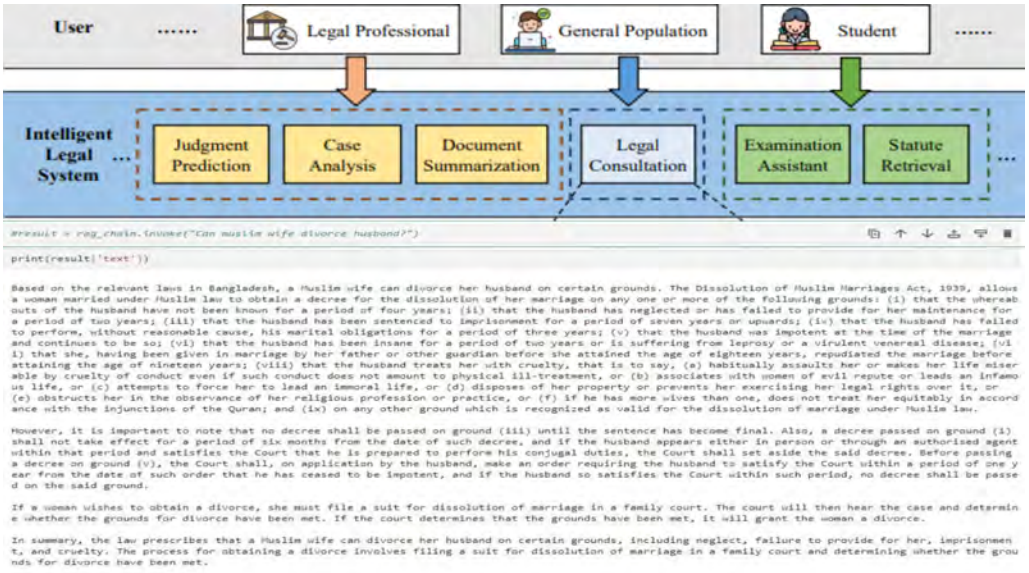


Figure 4.10: Model Input & Output.

Chapter 5

Human In The Loop

Our proposed system Human-in-the-loop shown in Figure:[5.1] includes a human certification system by hiring legal experts to evaluate the accuracy and relevancy. The model takes the queries from the users and generates answers with relevant sources and laws. But before sending it to the users, those answers will go through legal experts to check if the answer is correct or not shown in Figure:[5.2]. They will carefully test the answers and will improve it if necessary. And then the model will also learn from it for future reference. This process makes the information more accurate and reliable.

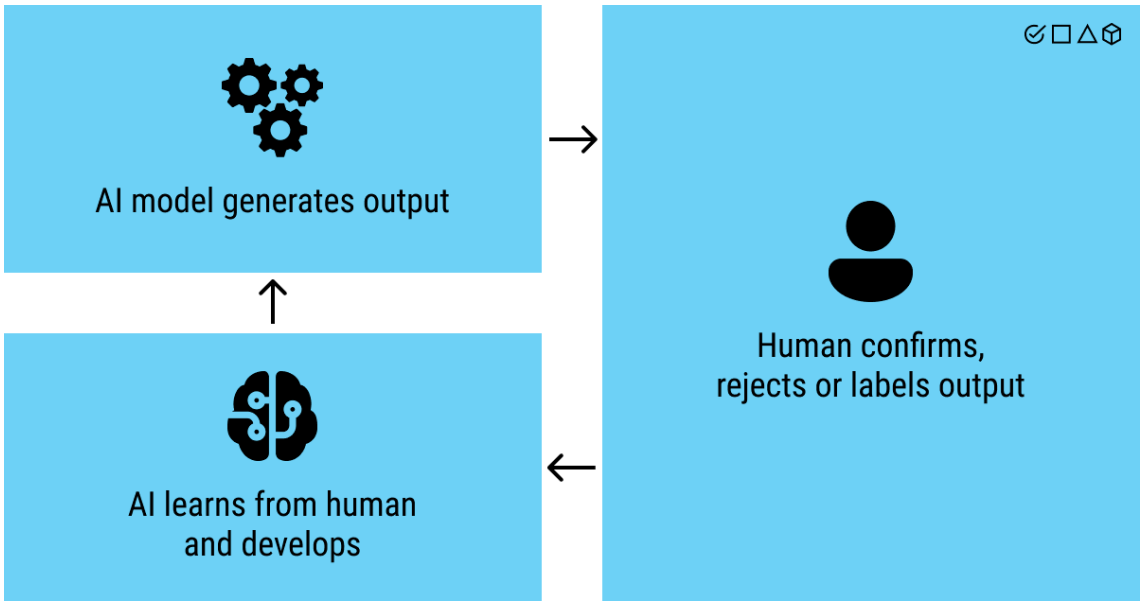


Figure 5.1: Human in the loop (HIL)

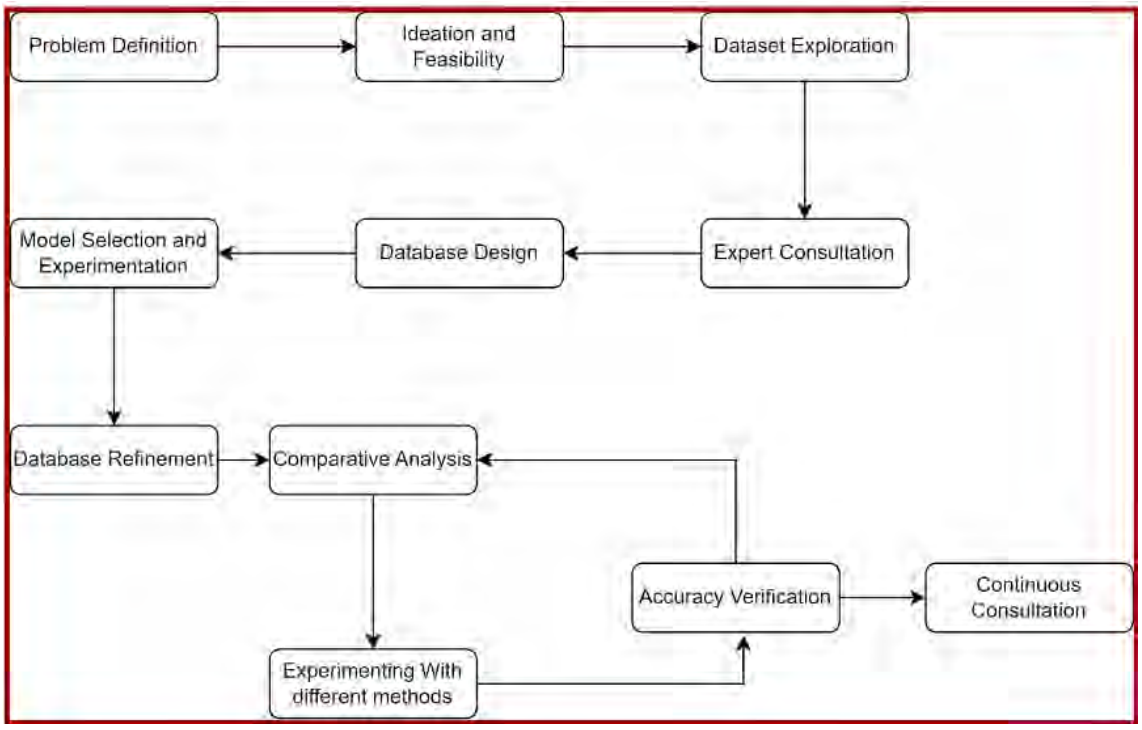


Figure 5.2: Pipeline after HIL

Chapter 6

Result Analysis

6.1 Results

In this research we have done two kinds of result analysis for evaluating our model's performance: i. Quantitative analysis and ii. Qualitative analysis.

6.2 Quantitative Analysis

6.2.1 Mass People Ratings

One of the important reasons behind our RAG based embedding system is that we wanted to make law accessible through easier methods as Delving into legal documents is difficult for the broader users. Thus we wanted to perform a readability test. For that reason we conducted an informal public survey in which we received 120 questionnaires from 12 people where each person asked 10 questions in a Google form then using the contact information from that same form we asked the participants to rate the answers on a scale of 1 to 10 based on their readability. A histogram shown in Figure:[6.1] of the distribution of ratings performed by the participants:

Figure 6.1: Readability Scores by participants.

From the histogram, we can see that the average readability score is 8.03. Depending on the score we classified the score into two categories:

- **Category A:** Readability between 6.0 and 8.0: There were 69 questions that were rated between 6.0 and 8.0. An example is shown in Figure:[6.2]

  Query: Q purchased at auction an immovable property belonging to a co-operative society at Tk. ninety lac. It was proved through evidence that X, secretary of the said co-operative society, in collusion with Y and Z, forged a document in favor of Z mentioning therein Tk. forty-five lac as the value of the property. Under which sections of the Penal Code, 1860 X, Y, and Z may be found guilty and sentenced? Answer with reference to relevant provisions.

  The principle that "pleading shall contain only material facts" is a fundamental rule under Order VI of the Code of Civil Procedure, 1908. As discussed in the given documents, this rule requires that pleadings should only include essential facts relevant to establishing the plaintiff's cause of action or the defendant's defense, without unnecessary details, arguments, or evidence. The purpose is to focus the case on the real issues in dispute between parties. As noted in the chunks, Order VI Rule 7 prohibits raising new grounds of claim or allegations inconsistent with previous pleadings except by amendment. The documents also highlight that courts have emphasized the importance of pleading only material facts to determine the real issues between parties, and have discouraged unnecessary details or argumentative pleadings. Failure to comply with this rule may result in the court ordering better particulars to be furnished or allowing amendments to include material facts initially omitted, as per Order VI Rule 17.
  One glimpse of our average readability score generated

  One of the reasons we think for such a low readability score we got on this paragraph is because it contains a lot of academic wording and law sections as references.

  Figure 6.2: Example answer for category-A.

  After using the TextStat python library to analyze the lowly generated test in Figure:[6.3] we can see:

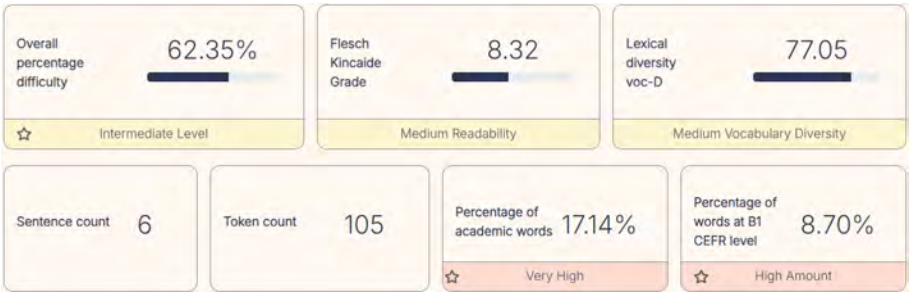


  Figure 6.3: Readability test for category-A.

- **Category B:** Readability between 9.0 to 10: There were 69 questions that were rated between 9.0 to 10.An example is shown in Figure:[6.4]

  Even though this query has been highly scored shown in Figure:[6.5] the Text-Stat shows medium readability with Very High academic wording.

Query: What is the law regarding public smoking?

The law regarding public smoking in Bangladesh is regulated by the Smoking and Tobacco Products Usage Control Act of 2005. Section 4 of the Act prohibits smoking in public places, and any person who commits this offense shall be punished with a fine that may extend to Taka 300. Public places include educational institutions, government offices, hospitals and clinics, court buildings, railway stations, public toilets, children's parks, fairs or passenger shelters, any other place declared smoking-free by the government or local government, and public transport means. The law also seeks to regulate the sale and distribution of tobacco products. The Smoking and Tobacco Products Usage Control Rules 2015 provide for measures to control smoking in public places, including the designation of smoking areas. The law also allows for the expulsion of a person who is smoking in a public place or on public transport by an authorized official and provides for the possibility of making a written complaint in the prescribed manner to the court for trial.

Figure 6.4: Example answer for category-B.

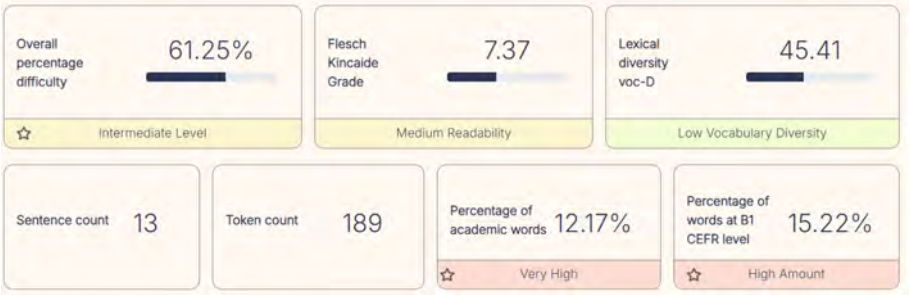


Figure 6.5: Readability test for category-B.

6.2.2 Overall Readability Score

The overall readability score is shown in Figure:[6.6] using Textstat library from Python on ten answers to the query using llama 3.1 8b instruct model:
The survey results demonstrate that our model performs well in readability in test scores. This performance is promising but the test result from the Textstat library shows otherwise. The reason will be discussed in the error analysis section.

6.2.3 Expert Evaluation

For further testing our model we have involved a legal expert to evaluate. First, we asked the expert to give us some queries. We generated answers to those queries and sent them back. We asked him to rate the questions out of 0 to 100. After getting the results we saw that the average score was 69.3 shown in Figure:[6.7]. The highest score given by the expert was 82 and the lowest was 52. Here question number 2,5,6,7 score was above average and questions number 3 and 4 showed that we need to improve on those areas.

6.2.4 Legal BERTScore

While BERTScore provides valuable insights into the precision, recall, and F1-score based on contextual embeddings, it has certain disadvantages. Specifically, BERTScore does not account for the hallucinations of domain-specific text, such as

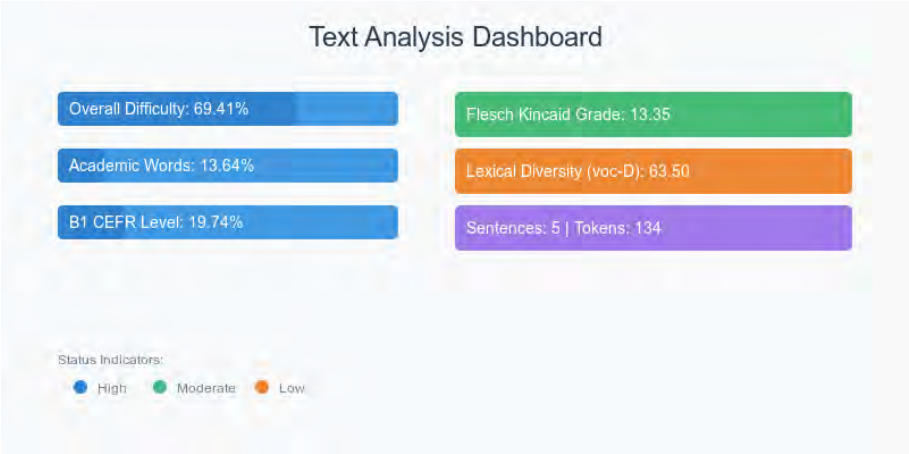


Figure 6.6: Overall readability score.

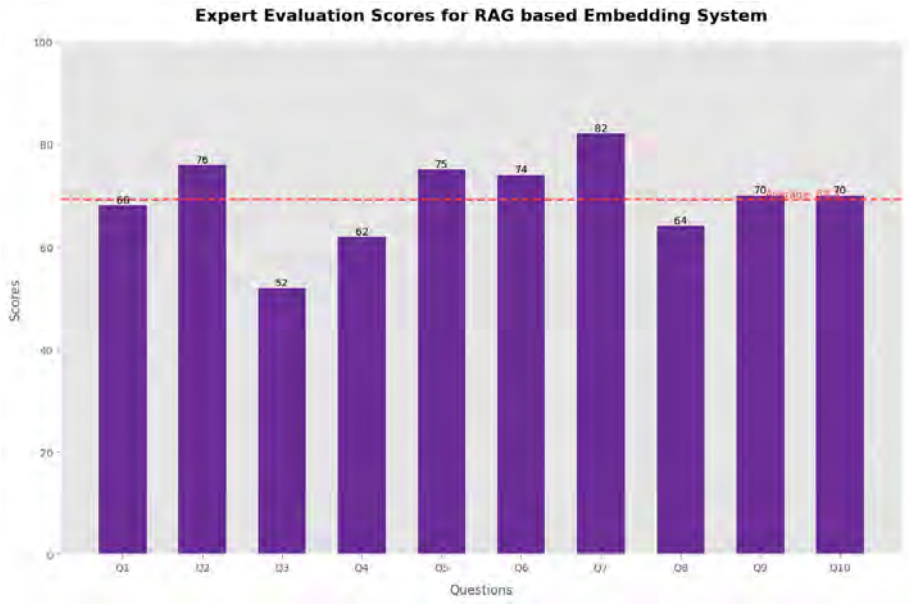


Figure 6.7: Expert evaluation scores.

legal language, where subtle differences in terminology and phrasing can significantly impact the accuracy of answers. This is why we opted for **Legal-BERTScore**, a domain-adapted version of BERT, which is more suited to tasks involving legal texts. Our metric utilizes two primary scores to evaluate the performance of models.
The first score is an F1 score. Which is measured using the semantic similarity between the generated answers and the correct answers using cosine similarity of embeddings produced by a legal-specific BERT model called legal-bert-base-uncased. [34]The difference between the F1 score produced by bertscore is that it compares exact token by token which is good for fine-grained information context similarities checking. It is not a suitable metric for domain-specific evaluations. In the legal domain, we already know that both the query and generated response will contain domain-specific terminology and concepts. Our approach operates at a higher semantic level by computing cosine similarities between entire sentences. This approach is suitable for us, where understanding the overall semantic meaning and relationship between sentences is more important than exact token matches.

The second score termed the hallucinations score, is a composite measure that combines assessments of misinformation, unnecessary sentences, and entity-based hallucinations.
By employing these two scores Shown in Table:[6.1], LegalBertScore provides a more comprehensive evaluation of both the semantic correctness and the factual reliability of generated legal responses. Legal-BERT is fine-tuned for understanding the unique semantics and structure of legal documents such that we feel obligated to take LegalBertScore into account in our overall research analysis.

Model	**F1 Score**	**Hallucination Score**
ChatGPT-3.5	0.625	0.617023
ChatGPT-4	1.000	0.649219
Claude 3.5	0.875	0.585438
LLama 3.1 8b Instruct RAG	0.875	0.500484
LLama 3.1 8b Instruct	0.625	0.603831
Mistral RAG	0.750	0.570872

Table 6.1: F1 Score and Hallucination Score Comparison

To further evaluate the model we used bar at law MCQ question 2023 to test our hypothesis shown in Table:[6.2].

Model	**Total Questions**	**Correct**	**Incorrect**	**Accuracy**
ChatGPT	100	64	36	64%
Claude	100	54	46	54%
Mistral	100	47	53	47%
Llama	100	68	32	68%

Table 6.2: AI Model Performance Comparison

The performance of AI models on the Bar at Law MCQ 2023 exam reveals important insights into the current state of legal AI. Firstly we would like to emphasize that both Chatgpt 4 and Claude 3.5 are larger models with 1.7 trillion parameters and 175 billion parameters respectively. Although parameter size does not equate to performance, a larger context window and better reasoning help in any generalizing tasks. That's why these results highlight that specialized training and domain-specific knowledge, rather than sheer model size, are crucial for legal reasoning tasks. The strong performance of the RAG-enhanced LLama model suggests that domain specific tasks such as this nature require the LLM and continuous finetuning with retrieval mechanisms for legal applications. As AI approaches human-level performance on such standardized tests, it sets new benchmarks for legal AI capabilities while also revealing areas for further improvement in accuracy and reasoning within the legal domain.

6.3 Qualitative Results

Upon our research, we found that Evaluating RAG-based legal question-answering systems using generative AI is very challenging. Traditional metrics like ROUGE

and BLEU scores fall short because they primarily focus on surface-level text similarity and exact word matches, which isn't suitable for legal content. These scores were made to test the quality of text summarization and machine translation. Both heavily focus on ngrams formation which is not suitable for LLM based tasks. Legal answers often require precise interpretation of laws, accurate citation of precedents, and complex reasoning chains that can be expressed in multiple yet valid ways. Two completely different phrasings could both be legally correct answers, but metrics like ROUGE and BLEU scores would score them differently.
This makes it necessary to often rely on expert human evaluation or develop specialized metrics that can account for legal domain-specific requirements, such as citation accuracy, reasoning validity, and proper application of legal principles. However, because of resource limitations and lack of proper funding the Human evaluation part was difficult to be constructed at a satisfactory level. So in order to do a proper evaluation, we found BERTScore. Unlike traditional metrics like BLEU or ROUGE, which rely on exact n-gram matching, BERTScore leverages semantic similarity, meaning it can account for the contextual meaning of word by word basis. [5] While BERTScore provides valuable insights into the precision, recall, and F1-score based on contextual embeddings, it has serious disadvantages. Specifically, BERTScore does not account for domain specific text where subtle subtle differences in terminology and phrasing impacts the accuracy of answers. [10]Therefore, we decided to go for new evaluation metrics that are more suited to tasks involving legal texts and also lack the overall disadvantage of BERTScore. For an overview, the BERTScore is shown in Table:[6.3].

Model	**Precision**	**Recall**	**F1**
ChatGPT-3.5	0.8514	0.8246	0.8377
ChatGPT-4	0.9003	0.8751	0.8875
Claude 3.5	0.8398	0.8684	0.8538
LLama 3.1 8b Instruct RAG	0.8932	0.8635	0.8780
LLama	0.7876	0.8408	0.8132
Mistral 7b instruct v0.3 RAG	0.6365	0.8027	0.7094

Table 6.3: BERTScore Results for Various Models

BERTScore provided valuable insights through precision, recall, and F1 scores based on contextual embeddings. However, it too presented limitations, particularly in accounting for the subtle differences in terminology and phrasing that significantly impact answer accuracy in domain-specific legal texts. For example one of the questions: **How are divorces, dissolution of marriages, or judicial separations handled for Christians, Hindus, Buddhists, or people of other religions except for Muslims in Bangladesh?**, Chatgpt4 mentions in Figure:[6.8].
Despite it being not a real law according to the government of the People's Republic of Bangladesh. In another portion of the answer, the model is getting the act correctly but explaining it wrongly shown in Figure:[6.9].
The statement chatGPT4 provided is partially accurate but the main contradiction lies in how divorce grounds are answered in the prompt. A similar problem also is present in Claude 3.5 sonnet and in chatGPT3.5 also in llama instruct
Finally despite being specifically mentioned in the query answers for Christians, Hindus, Buddhists, or people of other religions except for Muslims in Bangladesh

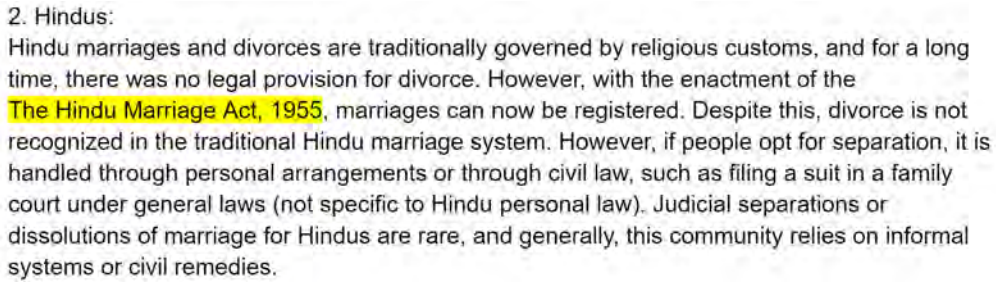


Figure 6.8: First portion of the answer.

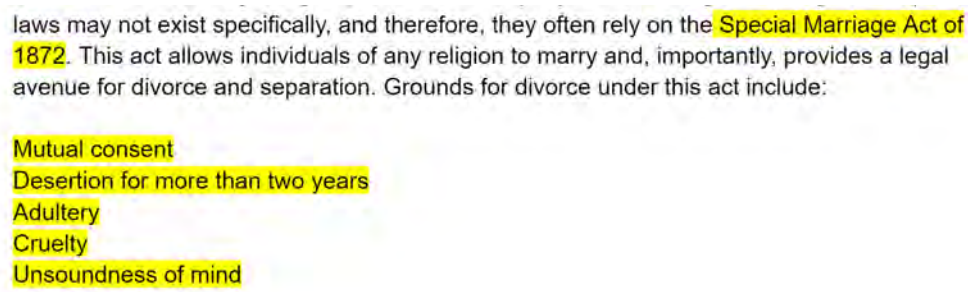


Figure 6.9: Second portion of the answer.

both chatGPTs, Claude, and even mistral for some reason mentioned laws from India instead of Bangladesh.Answer generated by Claude 3.5 sonnet shown in Figure[6.10]:

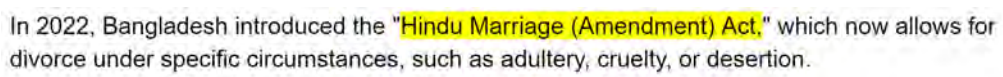


Figure 6.10: Second portion of the answer.

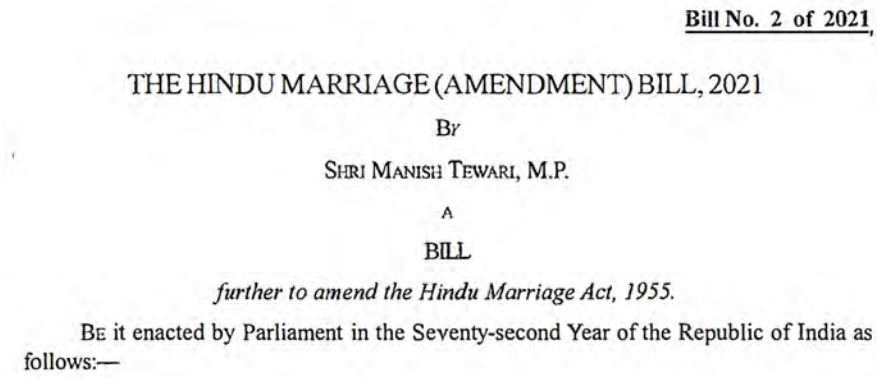
The act is not clear or accurate. Despite searching for this information. It was not available on the ”bdlaws” website. The only reference we could find to the existence of the act is shown in Figure[6.11]:
Then we asked our proposed Llama Instruct model to give answers to the same questions shown in Figure:[6.12]. It gave a detailed and relevant answer.
Again, we asked our proposed Mistral RAG model the same question, and the answer is shown in Figure:[6.13]. It gave a detailed and relevant answer.
Thus after these obvious blunders showcased by non-RAG models, the limitations of using BERTScore as a sole metric for performance evaluation were obvious.

6.3.1 Comparative analysis of the performance of different models on recent events:

We asked different AI models a question based on recent topics in Bangladesh to see how relevant the answer is. The question was:

Question: What is the legal validity of the current Interim government?

First, we asked this question to ChatGPT(the response is shown in Figure:[6.14]) and Claude sonnet(the response is shown in Figure:[6.15]). But the result was not



Bill No. 2 of 2021

THE HINDU MARRIAGE (AMENDMENT) BILL, 2021

BY

SHRI MANISH TEWARI, M.P.

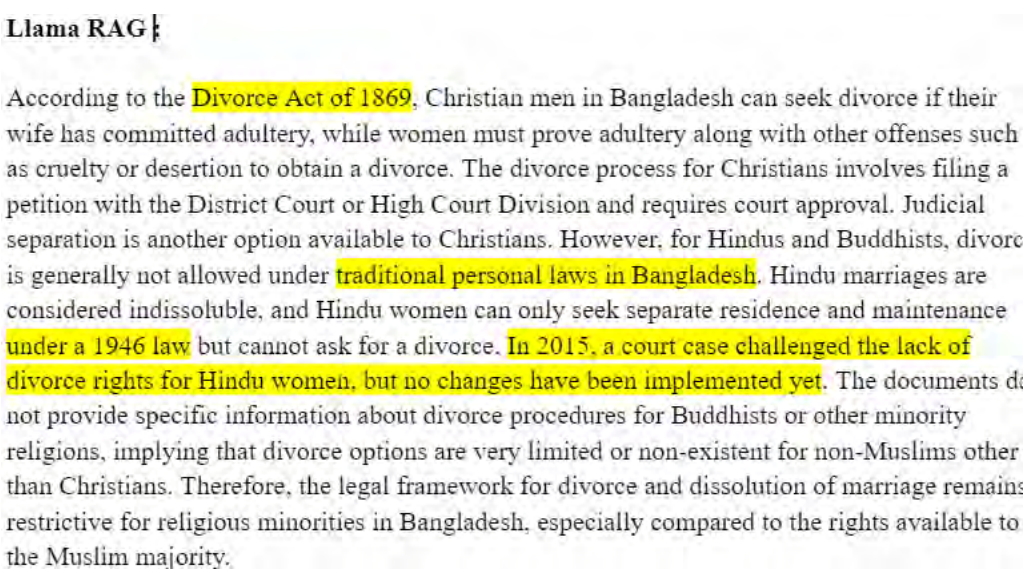
A

BILL

further to amend the Hindu Marriage Act, 1955.

BE it enacted by Parliament in the Seventy-second Year of the Republic of India as follows:—

Figure 6.11: Reference of the Hindu Marriage Act.



Llama RAG

According to the Divorce Act of 1869, Christian men in Bangladesh can seek divorce if their wife has committed adultery, while women must prove adultery along with other offenses such as cruelty or desertion to obtain a divorce. The divorce process for Christians involves filing a petition with the District Court or High Court Division and requires court approval. Judicial separation is another option available to Christians. However, for Hindus and Buddhists, divorc is generally not allowed under traditional personal laws in Bangladesh. Hindu marriages are considered indissoluble, and Hindu women can only seek separate residence and maintenance under a 1946 law but cannot ask for a divorce. In 2015, a court case challenged the lack of divorce rights for Hindu women, but no changes have been implemented yet. The documents d not provide specific information about divorce procedures for Buddhists or other minority religions, implying that divorce options are very limited or non-existent for non-Muslims other than Christians. Therefore, the legal framework for divorce and dissolution of marriage remains restrictive for religious minorities in Bangladesh, especially compared to the rights available to the Muslim majority.

Figure 6.12: Generated answer from proposed Llama RAG model.

that impressive; they gave very vague answers. Both models gave these answers after 3 shots.
Then we asked our proposed Llama Instruct model to give answers to the same questions shown in Figure:[6.16]. It gave a detailed and relevant answer.
We can clearly see that despite being a large model ChatGPT and Claude sonnet fails to match the performance of our proposed model.

6.3.2 Insights about the model's performance

Through our results, we got some important insights about the performance of these models for legal question-answering tasks. Specifically, for the Bar at Law descriptive questions of 2023.

- **Limitations of Large, General-Purpose Models:** Even after having an impressive number of 1.8 trillion parameters and a large context window, ChatGPT-4 lacks in providing accurate and reliable answers to specific legal questions. Such limitations indicate that model size and general language understanding on their own do not ensure domain-specific expertise.

Mistral RAG :

For Christians, divorce is allowed but only on very limited grounds under the Divorce Act of 1869. Christian men can divorce if they allege adultery by the wife, while women have to prove adultery along with other offenses like cruelty or desertion. The divorce process for Christians involves filing a petition with the District Court or High Court Division and requires court approval. Judicial separation is also an option for Christians. For Hindus and Buddhists, divorce is generally not allowed under traditional personal laws in Bangladesh. Hindu marriages are considered indissoluble, and there is no provision for divorce. Hindu women can seek separate residence and maintenance under a 1946 law, but cannot ask for divorce. In 2015, a court case challenged the lack of divorce rights for Hindu women, but no changes have been implemented yet. The documents do not provide specific information about divorce procedures for Buddhists or other minority religions, but imply that divorce options are very limited or non-existent for non-Muslims other than Christians. Overall, the legal framework for divorce and dissolution of marriage remains restrictive for religious minorities in Bangladesh, especially compared to the rights available to the Muslim majority.

Figure 6.13: Generated answer from proposed Mistral RAG model.

ChatGPT (After 3 shots)

As of now, Bangladesh does not have a legal provision for an interim or caretaker government, as the system was abolished in 2011 through the 15th amendment to the Constitution. Previously, a non-partisan caretaker government oversaw elections, but now elections are held under the ruling government with oversight from the Election Commission. Opposition parties, especially the BNP, have called for the reinstatement of a neutral interim government to ensure fair elections, but this demand has not been met. Any move to establish an interim government would require a constitutional amendment or broad political consensus to be legally valid.

Figure 6.14: ChatGPT(After 3 shots).

- **Limitations of Large, General-Purpose Models:** Another powerful LLM, Claude 3.5 shows its inability to make consistent distinctions when it comes to Bangladeshi law vs other legal systems. This inability points to the critical need for models to have accurate and jurisdiction-specific legal knowledge. As legal knowledge does not have any scope to be a bit slightly wrong, so AI applications can not mix up different legal systems eventually leading to serious consequences.
- **Limitations of Large, General-Purpose Models:** With low hallucination scores and good bar exam performance, our proposed model LLama3.1-8b instruct RAG has the potential of Retrieval Augmented Generation in legal applications. With the help of nine thousand plus (9000+) legal documents, our model achieved a balance between general language understanding and specific legal knowledge.
- **Limitations of Large, General-Purpose Models:** There is a clear difference between ChatGPT-4's high BERTScore and its actual performance in answer quality and reliability. This difference highlights the limitations of purely semantic similarity-based matrices in the evaluation of legal question-answering systems. Ultimately suggesting the need for more comprehensive evaluation methods that take factual accuracy and legal relevance under con-

Claude 3.5 sonnet (After 3 shots)

The legal validity of Bangladesh's current interim government is a subject of significant controversy. Following a 2011 constitutional amendment that abolished the caretaker government system, the ruling Awami League party under Prime Minister Sheikh Hasina maintains an "election-time government" which they argue is constitutional. However, the opposition Bangladesh Nationalist Party (BNP) contests this arrangement, calling for a neutral caretaker government to oversee elections. This dispute reflects broader tensions in Bangladesh's political landscape, with concerns raised by opposition parties and some international observers about the fairness of the electoral process. The situation remains contentious, with ongoing debates about the legitimacy of the current governance structure in light of constitutional changes and political dynamics.

Figure 6.15: Claude 3.5 sonnet (After 3 shots).

LLama instruct model:

The legality of the interim government is justified based on both Kelsen's theory and the existing constitutional law of Bangladesh. According to Kelsen's theory, the legality of a government is determined by its ability to maintain effective control over the state's machinery. In the case of the interim government, the president, with the utmost responsibility of the state, swore in this interim government, giving it legality and enunciating it as a de jure government. Additionally, the effective control the interim government gained over state machinery, through the controlling of civil servants, law enforcement agencies, and the emphatic and visible support of the people at large, proves this interim government to be the de facto government as well. Furthermore, the recognition of the international community, including the United Nations and various countries, and their willingness to work with this new government is also a very important characteristic of a valid government of a country in this era of modern international law. The formation of the interim government led by Dr Muhammad Yunus was overseen by the president, which goes in line with the aforementioned constitutional law principle, since the vacuum created by the absence of the prime minister altogether elevates the duties of the president, thus instating this government as legal and de jure.

Figure 6.16: LLama instruct model.

sideration.

- **Limitations of Large, General-Purpose Models:** The competitive performance of our proposed model LLama 3.1 8b Instruct RAG compared to much larger models like ChatGPT-4 clearly shows that smaller models like ours when combined with domain-specific retrieval mechanisms, become highly effective in specialized domains like law in our case. On the other hand, mistral 7b instruct RAG is also sufficiently capable of answering the queries with minimal hallucination.

6.4 Error Analysis

- **Limited Context Window Leads to Insufficient Retrieval** One of the key challenges faced in our model is the limitation of the token size of our proposed model llama 8b instruct rag and mistral 7b instruct 128K tokens and 32K tokens respectively, but due to our V-ram limitations, we could not utilize more than 8k tokens. Unfortunately this results in insufficient retrieval of relevant documents. Due to this restriction, the model is unable to consider a broad enough range of contextually relevant documents, leading to

inaccuracies in the generated answers. When the model processes a query, it sometimes retrieves an incomplete or inadequate set of documents, which significantly impacts the correctness and completeness of the response. This can be mitigated by using a better GPU with a bigger V-ram.

- **Complex Vocabulary from Domain Adaptation** Previously we mentioned that our model had a higher readability score but performed less on the Textstat python library this is because after performing domain adaptation on our legal corpus the vocabulary of the model became more specialized, aligning with domain-specific jargon. While this improves performance in technical or legal contexts, it poses a problem for general users or those unfamiliar with the domain. The language generated by the model may become overly complex or difficult for non-experts to understand, reducing accessibility and usability for a broader audience.

- **Decreased Reasoning Ability for Scenario-Based Questions** As a LLM our model lacks in parameter size also compared to SOTA models our LLM lacks in reasoning performance. That's why many of our generated prompts lacked a deeper understanding of the overall context. While it performs well in direct or fact-based queries, its ability to understand and logically infer answers in situations requiring deeper reasoning or hypothetical scenarios, is limited. This can be mitigated by using Paid LLM services like chatGPT or using bigger parameter sized open source models like LLama 3.1 70B or LLama 3.1 405B models.

Chapter 7

Conclusion

This paper aims to advance the field of legal expertise through an AI-powered legal assistant, 'BanglaLawBot,' for Bangladesh, an innovative system that combines a standard LLM model with retrieval-augmented generation. Legal practice requires critical knowledge that is often inaccessible to the general public, and by addressing these challenges and tackling the language barrier, we have taken a significant step by creating three unique datasets, fine-tuning large language models, and integrating them with a robust RAG system. Key contributions of this research include the introduction of these datasets, the combination of RAG systems with powerful LLMs, and the development of the LegalBERT score. Despite the challenges, our model outperforms available LLMs like ChatGPT and Claude in various aspects, even with a smaller model size. We also introduced a Human-in-the-Loop system and validation processes involving public and legal experts, furthering the field while addressing societal needs. However, it still has several limitations. To explain, the primary challenge we faced is with resource constraints. Most of the legal data are copyrighted making it impossible to get access to high-quality datasets without proper funding. Not only datasets but also financial limitations prevented us from hiring a full-time expert for evaluation. This challenge was also mentioned by Zhong et al. (2020). Additionally, with our limited resources, the sample size of our expert evaluation was a collection of only ten questions. This small sample size might lead to bias while not covering all the legal domains. a limitation also identified by Chalkidis et al. (2020) due to the lack of large language datasets for non-English speakers. Again the evaluation of the small sample size was done by only one evaluator making it a concern. Another undeniable challenge we faced is with computational power limiting the use of better models and getting a better outcome from our proposed framework. To address these limitations and get better output with our framework we will look for funding to acquire high-quality, copyrighted legal corpora and scale up computational infrastructure to train larger models. Also, we plan to deploy our legal embedding system as an app and a dedicated website for the Legal RAG system. We also intend to fine-tune embedding models specifically for the legal domain and explore the use of state-of-the-art LLMs such as ChatGPT and Claude for enhanced performance. This work marks the beginning of a legal literacy revolution in Bangladesh, making legal documents and knowledge accessible to the broader population.

Bibliography

[1] Z. Nine and F. Hoque, "Promoting access to justice through legal aid in bangladesh: A critical analysis," *IOSR Journal of Humanities and Social Science*, vol. 24, no. 8, pp. 55–60, 2016, Series-6. [Online]. Available: https://www.iosrjournals.org/iosr-jhss/papers/Vol.%2024%20Issue8/Series-6/H2408065560.pdf.

[2] F. AKTER, "Legal aid for ensuring access to justice in bangladesh: A paradox?" *Asian Journal of Law and Society*, vol. 4, no. 1, 2017.

[3] R. Islam, "Access to justice through legal aid: A study in bangladesh," *American International Journal of Social Science Research*, vol. 1, no. 1, pp. 22–32, Aug. 2017. DOI: 10.46281/aijssr.v1i1.159. [Online]. Available: https://www.cribfb.com/journal/index.php/aijssr/article/view/159.

[4] M. Raju and F. Ahmed, *Access to what?* Accessed: 2024-10-16, 2019. [Online]. Available: https://www.researchgate.net/publication/330077674_Access_to_What.

[5] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[6] S. Gururangan, A. Marasović, S. Swayamdipta, *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[7] T. Hasan, A. Bhattacharjee, K. Samin, *et al.*, "Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2612–2623.

[8] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[9] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[10] M. Hanna and O. Bojar, "A fine-grained analysis of bertscore," in *Proceedings of the Sixth Conference on Machine Translation*, 2021, pp. 507–517.

[11] S. Khazaeli, J. Punuru, C. Morris, *et al.*, "A free format legal question answering system," in *Proceedings of the Natural Legal Language Processing Workshop 2021*, 2021, pp. 107–113.

[12] S. Spohr Maximilian de Souza, *Technology, Innovation and Access to Justice — library.oapen.org*, https://library.oapen.org/handle/20.500.12657/62325, 2021.

[13] M. Romel, "Present scenario of legal aid in bangladesh: An overview," *IOSR Journal of Humanities and Social Science*, vol. 27, pp. 01–07, Sep. 2022. DOI: 10.9790/0837-2709060107.

[14] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, 2023.

[15] A. Q. Jiang, A. Sablayrolles, A. Mensch, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[16] C. Krishna, "Prompt generate train (pgt): Few-shot domain adaption of retrieval augmented generation models for open book question-answering.," *CoRR*, 2023.

[17] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large language models in law: A survey," *arXiv preprint arXiv:2312.03718*, 2023.

[18] C. Ryu, S. Lee, S. Pang, *et al.*, "Retrieval-based evaluation for llms: A case study in korean legal qa," in *Proceedings of the Natural Legal Language Processing Workshop 2023*, 2023, pp. 132–137.

[19] J. Saad-Falcon, O. Khattab, K. Santhanam, *et al.*, "Udapdr: Unsupervised domain adaptation via llm prompting and distillation of rerankers," *arXiv preprint arXiv:2303.00807*, 2023.

[20] R. Shui, Y. Cao, X. Wang, and T.-S. Chua, "A comprehensive evaluation of large language models on legal judgment prediction," *arXiv preprint arXiv:2310.11761*, 2023.

[21] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.

[22] A. Dubey, A. Jauhri, A. Pandey, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[23] J. Eschbach-Dymanus, F. Essenberger, B. Buschbeck, and M. Exel, "Exploring the effectiveness of llm domain adaptation for business it machine translation," in *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, 2024, pp. 610–622.

[24] S. Kim, S. Choi, and M. Jeong, "Efficient and effective vocabulary expansion towards multilingual large language models," *arXiv preprint arXiv:2402.14714*, 2024.

[25] S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha, "A literature survey on open source large language models," in *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, 2024, pp. 133–143.

[26] C. Lee, R. Roy, M. Xu, *et al.*, "Nv-embed: Improved techniques for training llms as generalist embedding models," *arXiv preprint arXiv:2405.17428*, 2024.

[27] N. F. Liu, K. Lin, J. Hewitt, *et al.*, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

[28] L. Martin, N. Whitehouse, S. Yiu, L. Catterson, and R. Perera, "Better call gpt, comparing large language models against lawyers," *arXiv preprint arXiv:2401.16212*, 2024.

[29] D. McDonald, R. Papadopoulos, and L. Benningfield, "Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark," *Authorea Preprints*, 2024.

[30] N. Prasad, M. Boughanem, and T. Dkaki, "Exploring large language models and hierarchical frameworks for classification of large unstructured legal documents," in *European Conference on Information Retrieval*, Springer, 2024, pp. 221–237.

[31] S. Setty, K. Jijo, E. Chung, and N. Vidra, "Improving retrieval for rag based question answering models on financial documents," *arXiv preprint arXiv:2404.07221*, 2024.

[32] K. Shi, X. Sun, Q. Li, and G. Xu, "Compressing long context for enhancing rag with amr-based concept distillation," *arXiv preprint arXiv:2405.03085*, 2024.

[33] S. Siriwardhana, M. McQuade, T. Gauthier, *et al.*, "Domain adaptation of llama3-70b-instruct through continual pre-training and model merging: A comprehensive evaluation," *arXiv preprint arXiv:2406.14971*, 2024.

[34] H. Steck, C. Ekanadham, and N. Kallus, "Is cosine-similarity of embeddings really about similarity?" In *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 887–890.

[35] X. Wang, J. Sun, and C. Qi, "Ceda-tqa: Context enhancement and domain adaptation method for textbook qa based on llm and rag," in *2024 International Conference on Networking and Network Applications (NaNA)*, IEEE, 2024, pp. 263–268.

[36] M. Xu, W. Yin, D. Cai, *et al.*, "A survey of resource-efficient llm and multimodal foundation models," *arXiv preprint arXiv:2401.08092*, 2024.

[37] L. Ye, Z. Lei, J. Yin, Q. Chen, J. Zhou, and L. He, "Boosting conversational question answering with fine-grained retrieval-augmentation and self-check," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2301–2305.