Exploring the Limitations of AI in Legal Reasoning: Challenges, Constraints, and Implications for Bangladeshi Law

by

Md. Raiyan Uddin
21301613
Onik Deb Nath Partha
22301572
Samuzzal Swapno
24241327
Hafsa Obedy
21201805
Sayeeb Hossain
22101498

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 and Engineering

Department of Computer Science and Engineering
Brac University
January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing 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: Student's Full Name & Signature:

Md. Raiyan Uddin
21301613

Onik Deb Nath Partha
22301572

Samuzzal Swapno
24241327

Hafsa Obedy
21201805

Sayeeb Hossain
22101498

Approval

The thesis titled "Exploring the Limitations of AI in Legal Reasoning: Challenges, constraints, and implications for Bangladeshi Law " submitted by

1. Md. Raiyan Uddin (21301613)
2. Onik Deb Nath Partha (22301572)
3. Samuzzal Swapno (24241327)
4. Hafsa Obedy (21201805)
5. Sayeeb Hossain (22101498)

of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall, 2025.

Examining Committee:

Supervisor:
(Member)

Moin Mostakim
Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

A B M Muntasir Rahman
Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis 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
Department of Computer Science and Engineering
Brac University

Abstract

Artificial Intelligence specially Large Language Models, are significantly promising for automating legal tasks as they interpret texts really well. However, in the context of the Bangladeshi Legal system it is constrained by fundamental limitations in bias, contextual misalignment and accountability gap. The research is an empirical analysis on the possibilities of these LLMs in making statutory decisions, even with the presence of misinterpretations of these LLMs in Bangladeshi context. It is based on a primary dataset on Bangladeshi legal criminal cases, that compares multiple architectures for predicting legal sections and generating analogies and finding out the need for rigorous validation, oversight, and domain specific adaptation to increase AI's applicability in Bangladeshi legal systems. Our findigs show that comparing multiple variations of sequence, transformer and state space architectures, it was concluded that metrics like F1 score, recall in section prediction and Cosine-similarity, BERTScore, Jaccard Similarity in analogy generation prevailed. This research proposes that the usage of these models in a proper section prediction and analogy generation tasks is possible only if the limitatioms of such LLMs are highly taken into consideration. Creating an equilibrium between the results of such models and the limitations is essential to ensure that AI-driven legal reasoning remains fair, accurate, and accountable.

Keywords: Artificial Intelligence, Large Language Models, limitations, Bangladeshi Law, Legal Provision, Bias, Accountability Gap, Contextual Misalignment, Reasoning Errors, Validation, Domain specific Adaptation, Fairness, Accuracy, Accountability.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the last few years, the incorporation of Artificial Intelligence (AI), specially Large Language Models (LLMs), in the legal sector has been expanding quickly [20]. All these models are capable of summarizing legal documents, predicting the judgement, and legal research. Therefore, instead of having these advancements, applying LLM in the legal system is still full of various limitations, particularly in developing countries like Bangladesh where legal infrastructure are still being digitized and most of the places are still working manually which are still evolving.

A basic challenge in the complicated legal language and reasoning is that legal systems often work with domain specific terms which are contextually sensitive [46]. Moreover, the fundamental problem of generally used models while decisionn-making such as GPT, LLaMa or Transformers with a maximum of 1.76 trillion, 2 trillion and 1 trillion parameters respectively, is the "blackbox" nature of algorithms, that is, the decisions they conclude on are not interpretable [56], [30], [28]. Each parameter contributes to the learning of the model and such extensive layers of learning are impossible for humans to understand. As this is a very concerning issue for AI advancement, for example, tools like COMPAS are criticised for using racial injustice without proper accountability [56]. Similarly, in Europe, the SyRI welfare fraud detection system was ruled unlawful due to violating human rights and using biased methodology [56]. Other tools like HART (Harm Assessment Risk Tool), deployed in the UK or PredPol (Predictive Policing System), deployed in the USA, have their own high risk in usage because of the same "blackbox" issue amongst the models [16], [47].

Recent years have also seen issues with AI hallucinations in legal contexts, for instance, in Mata v. Avianca, Inc., U.S. lawyers were sanctioned after submitting fabricated legal citations produced by ChatGPT [32]. These incidents highlight the unreliability of AI tools when applied in sensitive legal environments. Furthermore, even advanced legal AI systems still generate false or misleading legal references in nearly one-fifth of cases according to a 2024 study on AI legal research tools.

In the context of Bangladesh, where legal interpretation mostly depends on human decision, and also because of cultural context, the limitations of AI have become more and more noticeable.The judiciary system suffers from backlogs since cases often go pend-

ing without judgement as well as there is a lack of digital infrastructure [18]. Usage of AI in such areas without dealing with these problems will lead to further complications. Additionally, local studies and reports have pointed out emerging problems related to AI-driven misinformation, deepfake content, and ambiguities in cyber law, where victims face challenges due to limited technical and legal frameworks.

Furthermore, some of the legal AI tools show promising outcomes but most of them are under development and being trained with datasets and jurisdictions which are not similar in context of Bangladesh's legal system. For that, these models often fail to deliver according to our nation and legal-cultural context [46]. These limitations indicate the need of assessment in shaping the AI-driven legal reasoning in the Bangladeshi context.

1.2 Rational of the Study or Motivation

This research paper is inspired by the uprising of using AI in legal systems globally, but lacking in Bangladesh. While strong economies are using large language models for tasks like legal cases and contract review, the Bangladeshi legal ecosystem is not yet developed to adopt these advancements [44]. One of the significant motivating factors is the presence of bias existing in AI models. The results of the LLM that is trained on the data of the developed countries can vary and provide potentially discriminatory results in legal reasoning [44]. This biasness can be consistent even after training the models in Bangladeshi context as they are pretrained in western legal context. The same has been raised in global issues, in research papers, in which the researchers discovered that algorithmic bias can be quite persistent even in controlled settings, because the structure of black-box models renders the process of identifying bias challenging and imposing restrictions on it challenging [51]. The prominent difference in cultural and legislative systems plays a huge role in this issue as well.

In addition, recent Bangladeshi research has highlighted that the country's justice system still faces significant digital gaps, data scarcity, and infrastructural limitations that restrict the adoption of AI in courts and legal analytics. The study argues that without proper localization of datasets, linguistic adaptation for Bangla legal texts, and digital case management, AI systems will continue to produce unreliable and contextually irrelevant outputs.

Their ethical use is further confused by the accountability gap which is especially evident when the decisions made by AI are not adequately justified or criticized [53]. Recently in literature, it was also determined that the explainability deficit of the AI legal systems reduces the fairness and public confidence in the justice. Further, contextual misalignment is where models do not affect the way legal matters are processed within a particular region, which presents numerous AI solutions that cannot be applied in real-life circumstances within the Bangladeshi law [44]. A study by Sun (2025) also emphasized that legal frameworks globally struggle to trace and assign responsibility for AI-driven legal decisions, especially when dynamic model updates alter reasoning paths.

Since legal decisions are capable enough to change life trajectory, the usage of AI without any error is much needed. Our research is to find the possbilities of using AI in legal cases,

courtroom or in judgements. The goal is not to demotivate AI usage rather to implement it in such a way that it is transparent, unbiased and free from any kind of misuse [15].

Within this study, we try to find the scopes of LLMs in Bangladeshi legal context,considering the issues mentioned beforehand. The idea of using these models and finding the possibilities in predicting legal decisions compared to that under the western system motivated the initiation of the research.

1.3 Problem Statement

The growing involvement of Artificial Intelligence (AI), specially Large Language Models (LLMs) in the legal fields worldwide has revolutionized the process of documenting legal texts or predictioning passed judgements. However, this advancement in technology involvement introduces severe constraints while implementing such legal ecosystem in a local region-specific environment for example, like that of Bangladesh. Even though all the existing models perform extremely well in understanding general language, they often face difficulties in correct interpretation of the contextual, linguistic and complex procedures that are unique to Bangladesh, more so because these models are highly trained on Western-centrin legal contexts [15].

The system of LegalRAG indicates that retrieval-augmented generation (RAG) can be required in managing multilingual or contextually diverse legal tasks because traditional LLMs are unable to handle cross-linguistic subtleties and do not correlate outputs with local law [15]. Furthermore, it has been noted that LLMs will hallucinate legal facts, misunderstand statutory provisions, and misunderstand the legal precedence between law and precedents and that is a major problem when used on a less represented legal system such as Bangladesh [18]. These misinterpretations can distort the very foundation of legal reasoning.

In addition to these technical issues, bias and accountability gap still remains as the biggest issues of AI in legal reasoning. The data-driven biases in models may cause bias in the interpretation of laws and the decision to legalize, and the lack of a specific person responsible for the decision-making of AI may represent both ethical and legal challenges. This sense of irresponsibility will make people doubt the final rulings and lose trust in the court system [20].

Although there is increasing interest in the application of AI to legal systems, less empirical research has been conducted on how AI models can perform specific legal reasoning tasks within the Bangladeshi legal system. Specifically, the literature assessing the viability of section prediction and legal analogy generation using AI and local legal data is not well concentrated. In order to fill this gap, the current study explores the feasibility of AI models to perform these two targeted tasks in order to evaluate their practical viability and identify any visible limitations in the context of Bangladeshi legal research, rather than attempting to address systemic or ethical issues on a larger scale.

1.4 Research Objectives

This research aims to explore the viable aspect of Artificial Intelligence (AI) application in the legal reasoning process of the Bangladeshi court system. Instead of focusing on theoretical validation, the research progresses from a conceptual review to an experimental investigation by developing and evaluating AI models on locally relevant legal data. This research is based on the manual construction of a primary dataset collected from Manupatra and practicing lawyers and the training of multiple task-specific models for section prediction and legal reasoning.

The study uses four different AI models where Bi-LSTM and Bi-GRU are used for section prediction, while T5 and Mamba are used for legal analogy generation. The purpose of the research, through these implementations, is to determine how effectively AI models can solve limited but significant legal reasoning problems in Bangladesh.

The specific objectives of this study are as follows:

- To construct a primary legal dataset on Bangladeshi context and form an experimental foundation for evaluating using AI-based legal reasoning models.
- To train and evaluate Bi-LSTM and Bi-GRU models to predict penal code sections of Bangladeshi legal law.
- To train and evaluate T5 and Mamba models to generate lawyer analogies that take after human legal reasoning based on statutory interpretation.
- To examine the common patterns of errors and weaknesses of AI model outputs, such as classifying parts of the law as anomalies and producing ineffective or contextually inaccurate legal analogies.
- To observe, theoritically, how issues such as contextual misalignment, accountability gap, and bias are present within AI-generated outputs.
- To offer task-oriented information about the opportunities and constraints of AI-assisted legal reasoning in Bangladesh, with an intention to inform future studies on the topic and responsible usage of AI in the legal field.

At the close of this study, the research aims to offer a well-founded and fact-based insight into how AI can be used as an aiding tool to legal reasoning in Bangladesh. Instead of providing solid answers to more general ethical or theoretical issues, the results provide practical viability, observed constraints, and ways of evolving the project further.

1.5 Scopes and Challenges

Scopes

This research explores the limitations of Artificial Intelligence (AI) in legal reasoning within the Bangladeshi legal context through an empirical and task-focused approach. The scope of the study is defined as follows:

1. Primary Dataset Creation and Validation:

The research develops a specialized legal dataset through manual collection from Manupatra via the MyAthens platform and the practicing legal lawyers. The curation of the dataset is done with the help of legal professionals to make it relevant and accurate and specifically focused on murder-related cases to preserve the context and control the complexity of the domain.

2. Architecture-Specific Model Implementation:

Rather than relying on generic large language models, the study uses task-specific architectures. Bi-LSTM and Bi-GRU models are used for section prediction, while T5 and Mamba models are used for legal analogy generation. This enables the empirical examination of the behaviour of different model architectures when applied to Bangladeshi legal reasoning tasks.

3. Task-Based and Exploratory Evaluation:

The evaluation focuses on task-specific performance, observed error patterns and qualitative analysis of generated outputs. The study focuses on the observation of model behavior and limitations in practice, rather than formal validation of theoretical frameworks or benchmarking against large-scale global legal AI systems.

Challenges

The research encountered several challenges during implementation, which influenced both model behavior and experimental scope:

1. Dataset Scarcity and Curation Complexity:

The main weakness of the study was that legal documents were to be collected and processed manually because there were few digital documents. It also involved further consultation with the lawyers to certify the materials of the cases gathered and to make sure that the data set is relevant, especially on the case of murder.

2. Architectural Sensitivity and Model Stability:

T5 and Mamba were applied to generation of legal analogies though it needed careful tuning to provide stable and contextually significant results. Architectural discrepancies in architectural behavior raised issues in consistency and reliability of the generative legal

reasoning tasks.

3. Code-Switching and Language Complexity:

The models showcase a likelihood to generalize or apply reasoning patterns taken from Western contexts. This required additional attention to contextual alignment during data preparation and analysis.

4. Contextual Overgeneralization Risk:

The models demonstrated a tendency to overgeneralize or apply reasoning patterns derived from Western legal contexts. This required additional attention to contextual alignment during data preparation and analysis.

Chapter 2

Literature Review

2.1 Preliminaries

In this study, the literature review is organized thematically to understand how Artificial Intelligence (AI) has been applied in the legal field and explore its potential and limitations in assisting the legal reasoning process. Research works that have been published between 2017 and 2025 are reviewed to include both foundational studies and recent developments. The literature review has been organized into themes, AI Accessibility and its Facilitation in Legal Systems, Limitations of AI in Legal Reasoning and Interpretability, Bias and importance of fairness in perspective of AI in legal reasoning, Accountability and Explainability Challenges, Contextual Misalignment in Diverse Legal Systems and Comparative Deep Learning Techniques in Legal Text Modelling and Their Relevance. This approach enables the discussion to be centered on the operational attributes and the constraints of AI in legal reasoning, as opposed to the theoretical attributes.

The rise of Artificial Intelligence and Large Languages Models (LLMs) in the legal field is the most discussed in the literature in the fields of document review, legal research and prediction of outcomes. Various studies confirm that although these models generate coherent and precise legal text and can replicate almost any legal document, they do learn legal rules and legal reasoning [20], [33], [15], [35]. This study focuses on the limitation of these models since legal reasoning is about the evaluation and analysis of the meaning of the provisions of the law and the relations between the facts, which in turn, stimulates the examination of specific functions such as predicting sections of statutes and generating analogies.[24], [41].

Previous literature mentions the most common challenges that arise when AI is used for legal reasoning. AI bias, opacity of the models, and AI/data mismatch are the most frequent challenges reported in the literature [36], [37], [39], [44]. This is especially true for Bangladesh which has a non-western and developing legal system, since the legal language, the legal system, and legal practice are fundamentally different from the ones in the majority of the AI training datasets [35], [38], [45]. In this thesis, these issues are considered as background concerns identified in earlier studies, rather than as problems that are directly modeled or measured.

Widely used evaluation methods for legal AI systems are also discussed in the literature, particularly in areas involving text classification and text generation. For statutory sec-

tion prediction, classification-based evaluation metrics are commonly used to assess how accurately models identify relevant legal provisions. For legal analogy generation, both semantic similarity measures such as cosine similarity, KL divergence, Bhattacharyya coefficient, and Jensen-Shannon divergence, and lexical overlap measures including Jaccard similarity, ROUGE, and BLEU, are typically employed to compare generated outputs with reference texts [12], [29], [57], [54], [52]. The selection of these evaluation techniques in the present study is based on their standard use in prior research and their comparability with existing literature.

2.2 Review of Existing Research

This section is a summary of the reviewed literature in the given document on the use of artificial intelligence (AI) in legal reasoning and specifically highlights its challenges as well as opportunities in the setting of the Bangladesh legal system. Five main thematic areas were recognized in the analysis and they include Accessibility and Implementation of AI in Legal Systems, Limitations of Legal Reasoning and Interpretability, Bias and Fairness in AI Uses, Accountability and Explainability Difficulties, and Contextual Misalignment in various Legal Landscapes. Those themes are mutually related closely. The subsections below explore these themes at length and take into account all the relevant studies present in the document explaining why the topic is relevant to the thesis of exploring the boundaries of AI in the legal system in Bangladesh [12]. The thematic discussions are organized in such a way that the mutual interrelations between them and consequences are stated within the framework of creating reliable, fair, and contextually correct legal domains of AI technologies.

2.2.1 AI Accessibility and its Facilitation in Legal Systems

The reviewed literature mostly highlighted the potential of AI and its probability in handling legal cases, more so in limited resource environments like that of Bangladeshi legal systems, where legal infrastructures are still underdeveloped and expert validations are questionable.

Wasi et al. (2024) trained a model on UKIL-DB-EN dataset named GPT2-UKIL EN, and did a thorough understanding of its deployment by using the Bangladeshi legal documentation and acts [48]. Their study addressed the lack of affordable legal support for the common folks and emphasized on creating a legal assistant that would be artificially intelligent and dedicated to ordinary citizens for legal support. Likewise, Hossain et al. (2024) ran the transformer-based Large Language Models (LLMs) for example LLaMA 3.1-8B and Mistral-7B through Retrieval-Augmented Generation (RAG) with the help of "Laws of Bangladesh" dataset, a public Q&A dataset and The Daily Star dataset each consisting of 1380, 2276 and 5633 entries respectively [38]. The work they did seeks the possibility of bridging the gap between the complexity of legal systems and public interpretation, focusing on accessing the natural language interfaces through AI. Kabir et al. (2025) emphasized more on this by running their own RAG experiment by developing a bilingual Q&A system on a large dataset such as Bangladesh Police Gazettes (2016-2023) and augmented it for AI compatibility [55]. Their used models were LLaMa 3.1 and Mixtral 8x7B with advanced RAG pipeline ended up with results as good as 0.82 cosine

similarity score, proving the fact that such models had improved AI precision even when resources were low. Nasir (2025) further enhances the positivity of these findings by doing rigorous running of AI tools like Kira, COMPASS and ROSS intelligence [53].

The results showed an automation of 22% legal cases which proves the increase in efficiency and the reduction in cost [53]. But these very studies show a significant drop in AI's transformative potential in such legal doctrines, particularly in Bangladesh that suffers from a lack of linguistic fluidity and proper resources. Having said that, these resources still hold the potential of accessibility to ordinary people, but their vulnerabilities in complex legal reasoning still prevails, more so in Bangladesh's unique legal framework. This takes the research to the next section, where the potential of accessibility is examined alongside bias, reasoning limitations, and accountability gaps to ensure proper confidence in using AI in legal systems.

2.2.2 Limitations of AI in Legal Reasoning and Interpretability

The literature indicates that AI has problems with complex legal reasoning. According to Wasi et al. (2024), GPT2-UKIL-EN is rather narrow in its context since it cannot perform complex legal tasks such as GPT-4, and subtle interpretations in Bangladesh are not an easy task [35]. Hossain et al. (2024) also discovered that their RAG-based models performed quite well in semantic matching and citing laws but not in real life-specific cases when it was required to analyze cases or apply logic to situations. This was manifested in poor Legal-BERTScore scores [36].

A number of LLMs, such as ChatGPT-4, ChatGPT-3.5, PaLM-2, and LLaMa-2, were trained on 15,000 U.S. federal court cases [35]. The models performed well in surface-level questions but with evident gaps in interpretability [34]. Fei et al. (2023) put 51 LLMs to the test with the LawBench benchmark in 51 legal tasks in Chinese. GPT-4 achieved the best overall score (52.35 out of 100) but scored lower on more challenging problems, such as article recitation and legal consultation, showing the weaknesses of even the best models in context-sensitive, multi-layered reasoning [24].

Choi and Schwarcz (2024) conducted experimental tests on GPT-4 with law school essay prompts and discovered that students occasionally scored poorly when using AI, particularly on tasks that required argumentation or critical thinking [33]. The same problems were found by Khazaeli et al. (2021), where their QA system performed well for fact-based questions but poorly for analytical ones due to jurisdictional differences in training data [21]. Kapoor et al. (2024) further note that interpretation of law involves judgment, abstraction, and policy reasoning, which AI systems struggle to handle due to their inability to comprehend social, cultural, or procedural variations [39].

Collectively, these studies indicate that existing AI solutions are not prepared to operate at the level of human legal expertise, particularly in a nation such as Bangladesh, where legislation comprises statutory ambiguity, religious authority, and colonial rules. Even advanced models face difficulties when reasoning requires deep contextual understanding.

2.2.3 Bias and importance of fairness in perspective of AI in legal reasoning

The matter of being biased in Ai represents itself as a critical obstruction to rely on legal problems. In the research done by Javed and Li (2024) focuses on linguistic bias in relevance to legal judiciary, there the core message shows that using biased language in judiciary can lead to changes in outcomes while using it in the training dataset of the model. Specifically, the model they have used, the SVM model can conclude the emotional, indicative, and stereotypical judgment with a great accuracy of 96.9 percent [39].

Hofmann et al. (2024) analyze dialectical prejudice and reaches to a conclusion that some LLM models like GPT-3.5 and GPT-4 showcases negative result towards those who speak in African-American English (AAE) are considered guilty more frequently more than 23% of the cases and 4.9% in the case of other language speakers like Chinese, Hindi, Arabic. This discrimination can also be applied in the context of different accent speakers in Bangladesh and also towards the tribal indigenous people who have their own language [37].

Guo et al. (2024) has represented a report in the context of biased outcomes where they break it down and reached the conclusion that since the model is heavily trained on western based dataset it has a biased behavior towards it and also external bias that emerges while being deployed of disparity [36]. Alvarez et al. (2024) in their NO BIAS project discusses in detail about different types of bias in AI which identifies as bias which already exists, the bias which occurs due to technical problems and the problems which occurs while AI is being used, and as a solution of this problem they have proposed an idea BML (bias management layer) that will resolve this matter and this is relevant to Bangladesh since we have a diverse cultural, linguistic system which has an impact on the legal system [34].

Another different kind of bias has been found out by Magesh et al. (2024) which is called bias through hallucination where AI tools can give wrong information or outcome and the percentage can be way up to 33 percent. In an experiment on 202 real legal questions Westlaw AI was given a 33 percent hallucination rate, In many cases, it failed to recognize the name of court cases and frequently failed to recognize legal authority levels [57].

Padiu et al. (2024) declares that the advanced models like LAwyer LLaMA and other like ChatLaw which are trained on such datasets as CAIL2018 (dataset from China) and ECHR (dataset of Europe), still show bias outcomes when solving problems that are out of the jurisdiction of the original datasets, which highlights that model must be trained with localized data to avoid such circumstances [46]. Dymrituk (2019) also showcased the issue of bias as a technical and ethical shortcoming about general LLM models such as GPT, which is not enough to develop legal systems in developing nations as these will cause chaos and misjudgments [18].

2.2.4 Accountability and Explainability Challenges

The lack of accountability and explainability of AI outputs are very pressing concerns in any legal system, particularly in countries such as Bangladesh, where the transparency of

the courts is low and the population does not trust them. The non-transparent nature of AI may also be a source of reputational loss when measures are not taken to protect it. According to Doshi-Velez et al. (2017), AI should allow stakeholders to explore automated reasoning by explaining in a local way what drives a particular decision [15].

The use of LLMs such as ChatGPT and Claude has been criticized in advanced jurisdictions. According to Waldon et al. (2025), three primary issues are recognized: the lack of clarity regarding responsibility in the case of errors, a lack of responsiveness to changes in the law, and unprovable results [50]. Court cases such as Snell and Deleon (2024) in America represent practical and ethical problems in judicial rulings based on AI [41].

Eben et al. (2025) provide a theory of transparent AI decision-making with reference to actual practice examples, such as the scandal in the UK Post Office, where automated systems resulted in false convictions because of a lack of control [51]. Dymrituk (2019) points out that it is morally necessary to hold someone or an institution responsible when AI takes lives, particularly in developing countries [18].

Nabben (2024) demonstrates that it is hard to distribute responsibility when more than one actor applies AI to the legal context. Her work underscores the necessity of having institutional clarity on the accountability of individuals who may produce AI recommendations that raise legal concerns, and the issue is particularly topical in Bangladesh [40].

These studies indicate that the absence of explainable and accountable AI endangers the legality of legal systems, citizens' trust, and justice. The literature agrees on the need for institutional structures and oversight mechanisms to protect against the erosion of core legal principles such as the entitlement to a fair hearing.

2.2.5 Contextual Misalignment in Diverse Legal Systems

The contextual misalignment is one of the biggest challenges facing AI in legal reasoning in the case of Bangladesh, with its mixed legal system, where statutory law and Muslim law are complemented by colonial law. Misalignment arises when AI models trained on datasets from other jurisdictions are implemented in a different legal, linguistic, and cultural setting. In Bangladesh, legal documents are dual-language and influenced by distinctive socio-cultural factors, which endanger the justice and reliability of AI-oriented systems. The literature also highlights that locally based information and human surveillance are crucial.

The problem is evident in studies of localized models. GPT2-UKIL-EN, which is trained on texts of Bangladeshi law, is not able to capture the qualitative nuances of law entirely or combine multiple sources of law [45]. RAG was applied to Bangladesh Police Gazettes (2016–2023) with a cosine similarity of 0.82 but showed lower performance (0.45–0.58) on out-of-context queries, indicating that even augmented models do not work effectively in Bangladesh's hybrid and multilingual system [48], [51].

According to a study by Magesh et al. (2024), the rate of hallucination in RAG-based tools per jurisdiction-specific question is 33%, leading to misinterpretation of legal au-

thorities [53]. Models trained on non-local data, therefore, do not capture the complexity of Bangladeshi law.

There are grave implications. Socol de la Osa and Remolina (2024) demonstrate that transparency and trust in AI products are not always enhanced by standardized products across borders due to the need to align them with local legal regulations [41]. According to Choi and Schwarcz (2024), the reliability of reasoning can only be ensured when grounding models are based on certain legal materials and embedded in context [33].

A number of solutions are suggested. To advance knowledge in a bilingual setting, Jiang et al. suggest relying on legal storytelling datasets [38]. Wasi, Kabir, and Magesh also emphasize the need to narrow down data that are geographically constrained and to tailor the models to specific judicial systems [45], [48], [53]. Human-in-the-loop validation is recommended to ensure correspondence between the output and local norms [37].

The contextual misalignment is thus a major obstacle to the adoption of AI in the legal system of Bangladesh, contributing to issues of bias and accountability. The results of these studies emphasize the necessity and relevance of localized datasets and human expertise to promote the accuracy and applicability of AI-based legal reasoning.

2.2.6 Comparative Deep Learning Techniques in Legal Text Modelling and Their Relevance

Over the past several years, researchers have been talking about the idea of using deep learning to better understand the understanding of law. A prominent method was introduced in which a Bi-LSTM was combined with fuzzy logic and metaphors recognition in order to read legal documents [15]. This design helps the model know the symbolic and indirect meanings that are common in legal writing. The Bi-LSTM network also utilizes both forward and backward processing, which helps to make the information more contextual. The model yielded a high macro-F1 score of 0.8005, indicating that there was high predictive accuracy, and the classification was consistent [15].

Cho et al. (2014) presented a similar function using Bi-GRU model with only two gate architectures: update and reset [20]. This design is less computationally intensive and faster than Bi-LSTM, with similar accuracy. With the addition of an attention layer, BiGRU is capable of identifying the most important parts of a piece of legal text, such as words that define an act or the outcome of a case.

Multilingual transformer applied to legal language processing in low-resource settings are mentioned in the literature as well. These models can be trained to complete different tasks with a text to text approach and then fine-tuned to achieve different objectives such analogy generation.

Medvedeva et al. highlighted that, legal texts are not as simple as general natural language because of their complex structure and symbolic usage of terms. Their study demonstrated that the traditional theories of language processing are unable to effectively decode legal text. Taking these into account, in our work we use Bi-LSTM and Bi-GRU models with attention mechanisms for classification of sections. These models

reflect both the flow of legal text forward and backward while highlighting key words. For analogy generation, the T5 transformer model was selected.

Considering these, our study utilizes Bi-LSTM and Bi-GRU models with attention mechanism for classification of the sections. These models capture the flow of text from forward to backward through the legal text and emphasize important words, a balance of understanding the text contextually and being able to implement the model computationally. For analogy generation, the T5 transformer model was used. T5 is scalable to any NLP task as it considers all of the NLP problems as text-to-text tasks, hence, invaluable analogies are generated in the legal space. Combined, these models are certain of context awareness, resource efficiency and suitability for bilingual legal documents.

2.2.7 Limitations of Existing Models and Justification of Selected Approaches

While many existing models such as CNNs, RNNs, and transformer-based models (BERT, RoBERTa) and hierarchical structures like the Hierarchical Matching Network (HMN) have shown promise in legal AI applications, they also reveal significant limitations that influenced our model selection. CNN models are fast and simple but analyze text in fixed-size blocks. Legal documents, however, rely heavily on relationships between sentences, clauses, and context. CNNs cannot follow long dependencies, which leads to loss of meaning and reduces interpretability in legal reasoning tasks.

RNN-based models initially addressed sequence problems but struggled with vanishing gradients when processing long documents, making it difficult to retain previous contextual information essential for case-based reasoning. HMN attempted to solve this with layered structures to match textual patterns at multiple levels. While effective for classifying specific crimes, HMN required complex preprocessing and feature engineering and was not easily adaptable to other legal domains or languages [19].

Transformer-based models such as BERT, Legal-BERT, and RoBERTa perform well in document understanding tasks but have practical drawbacks. They require substantial memory and GPU resources, which are often unavailable in low-resource environments like ours. They also depend on large labeled datasets for fine-tuning, which are scarce for Bangladeshi law.

Given these limitations, our research adopts Bi-LSTM and Bi-GRU with attention for classification and T5 for analogy generation. Bi-LSTM and Bi-GRU preserve long-range dependencies and process text bidirectionally, while the attention mechanism highlights key legal terms that determine a case's section or decision. The T5 transformer is more flexible and resource-friendly compared to large models like BERT, converting all tasks into a text-to-text format to effectively learn analogy generation from case descriptions and verdicts.

2.3 Summary of Key Findings

The analyzed literature demonstrates that the utilization of Artificial Intelligence (AI) in legal work has developed significantly over the last few years. Retrieval-Augmented Generation (RAG), BERT-based architectures and transformer-based language models have enhanced the efficiency in information retrieval and surface legal reasoning. These methods are specifically applicable in managing big legal corpora and bilingual issues that are applicable to legal systems like Bangladesh.

Nevertheless, significant limitations are always reported in the literature. Although the AI systems are more effective in structured or retrieval based tasks, they are unable to deal with legal problems that require reasoning such as statutory interpretation and situational analysis. The latter are more pronounced in low-resource legal settings, where the datasets in the domain are limited and the legal language is not the same as those of Western jurisdictions that are being trained on the models.

The common pitfalls of studies that are conducted are biased outputs, reduced transparency in AI-generated reasoning, and the inability to localize the models to local legal circumstances. The hallucinated legal references and superficial analogical reasoning are also likely to be generated by AI systems in particular in cognitively challenging tasks as well as in legally weak or misleading ones. Because of that, numerous researches focus on the necessity of human control and validation by experts.

On the whole, the literature indicates that AI may be used as an assistive tool in legal research and initial argumentation, yet its usefulness is disparate and case-specific. This indicates the necessity of the task-specific assessment based on the local legal data. To that end, the current study is dedicated to the study of AI performance and AI constraints in the statutory section prediction and generation of legal analogy in the Bangladeshi legal setting.

The following table summarizes the key findings from the reviewed studies, pointing out datasets, method/models, accuracy, and results:

Table 2.1: Summary of Key Findings from Reviewed Studies

Author and Year	**Focus Area**	**Method/Mode**	**Datasets**	**Results**
Magesh et al. (2024)	Legal Tool Reliability	Lexis+ AI, Westlaw AI, GPT-4	202 Real Legal Questions	Accuracy: 65% (Lexis+); Hallucination: 33% (Westlaw)
Javed and Li (2024)	Semantic Bias	SVM,Naive Bayes,MLP, KNN	CAIL (225,000 Judgments)	Accuracy: 96.9% (SVM); Bias Detection: Effective
Hofmann et al. (2024)	Dialect Bias	GPT-2,RoBERTa ,T5,GPT-3.5, GPT-4	Matched-Guise Sentence Pairs	Bias Rate: 23% Negative Traits; Conviction Bias: 4.9%

Author and Year	Focus Area	Method/Mode	Datasets	Results
Fei et al. (2023)	Legal Reasoning Benchmark	51 LLMs (GPT-4, ChatLaw)	CAIL2018, JEC-QA	Average Score: 52.35; Contextual Reasoning: 19.65
Kapoor et al. (2024)	Legal Applications	Review + Case Studies	Mixed (Global)	Accuracy Misleading: High ;Hallucination Risk: Significant
Wasi et al. (2024)	Legal Accessibility	GPT2-UKIL-EN	UKIL-DB-EN (Bangladesh Legal)	Qualitative Success; Contextual Misalignment: Notable; Reasoning Depth: Limited
Hossain et al. (2024)	Legal Assistance	RAG – LLaMA 3.1-8B, Mistral-7B	Laws of Bangladesh, Q&A, Judgments	Legal-BERT Score: High; Reasoning Depth: Weak
Guo et al. (2024)	Bias Analysis	Three-layer Framework	Mixed (Multilingual)	Bias Reduction: Moderate; Accuracy Trade-off: Noted
Doshi-Velez et al. (2017)	Accountability	Explainability Framework	Case Studies	Explainability Success: Variable; Transparency Gap: Persistent
Dahl et al. (2024)	Legal Hallucinations	ChatGPT-4, PaLM-2, LLaMA-2	15,000 US Federal Cases	Hallucination Rate: 58%-88%; Complex Task Failure: High
Socol de la Osa & Remolina (2024)	Judicial AI	Case Studies – Framework	Case studies from Colombia, Peru, and India	Contextual Misalignment: High; Efficiency Gain: Moderate
Kabir et al. (2025)	Bilingual QA	RAG + Llama 3.1	Bangladesh Police Gazettes	Cosine Similarity: 0.82; Contextual Tuning: Limited
Alvarez et al. (2024)	Fairness Policies	NoBIAS Architecture	EU Legal Data	Bias Mitigation: Effective; Localization Challenge: Significant
Jiang et al. (2024)	Legal Education	Storytelling + LLMs	LEGALSTORIES Dataset	Comprehension Gain: High; Analytical Limitation: Present
Khazaeli et al. (2021)	Legal QA	BM25 + BERT	LexisNexis + Natural Questions	F1 Score: 0.766; Contextual Coverage: Partial
Waldon et al. (2025)	Judicial AI	ChatGPT, Claude, Gemini	Real-time Monitoring	Reliability: Low; Prompt Sensitivity: High
Nasir (2025)	AI in Legal Systems	Review + Interviews	Mixed (Global Tools)	Efficiency Gain: 22%; Bias Concern: Critical

Author and Year	Focus Area	Method/Mode	Datasets	Results
Padiu et al. (2024)	LLMs on legal systems	LLMs (Lawyer LLaMA, ChatLaw, etc.), RAG, multi-agent systems	CAIL2018, ECHR, LLeQA	High precision, F1-scores; challenges with bias, contextual misalignment
Dymrituk (2019)	Ethical limits of LLMs in law	Qualitative Analysis	Legal documents, case law	Accuracy Low; Bias Concern: High
Choi and Schwarcz (2024)	Legal Education	GPT-4 (Grounded Prompts)	Law School Exams	Student Gain: 29% (simple); Complex Reasoning: Poor
Eben et al. (2025)	Ethical Liability	Theoretical Framework	Case Studies	Fairness Potential: High; Implementation Clarity: Lacking

This figure below is our proposed Workplan.

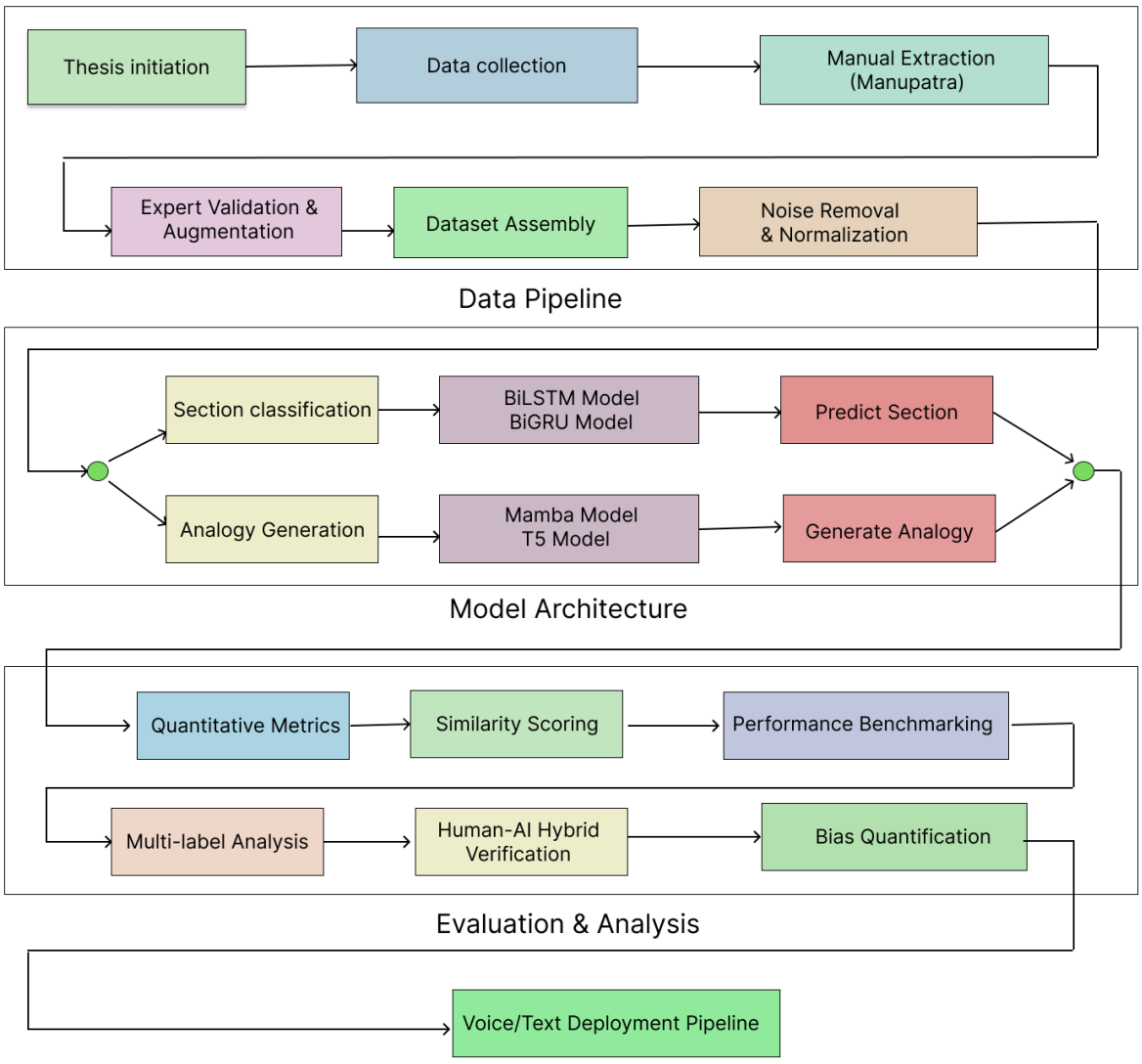


Figure 2.1: Proposed workflow

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specification and Requirement

The development and application of the legal reasoning model depend on the following specification and requirements:

3.1.1 Methodological Requirements

1. **Supervised Learning Paradigm:** We have used a causal language modeling approach where the model is fine-tuned on a dataset containing legal scenarios and verdicts. The primary targets are section prediction and legal analogy generation.

2. **Model Implementation:**

   - **Section Prediction:** Bi-directional Long Short-Term Memory (Bi-LSTM) and Bi-directional Gated Recurrent Unit (Bi-GRU) with attention mechanisms.
   - **Analogy Generation:** T5 Transformer model (220 million parameters) and Mamba State Space Model(130 million parameters).

3.1.2 Data Requirement

1. **Primary Dataset:** Case files contained in both Manupatra legal databases and private legal archives were used to develop a hybrid primary dataset. The data set will contain the case number, case file date, section law, scenario, case verdict and lawyer analogies.

2. **Preprocessing:** Regex-based cleaning was utilized to normalize legal terms and dates while preserving critical identifiers (e.g. case ID, section numbers). The dataset was split into 70% for training,15% for testing and 15% validation.

3.1.3 Minimum System Requirements: Software and Resources

To ensure the reproducibility of the legal reasoning model and analogy generation tasks, the following software and hardware environment is required:

- **Programming Language:** Python 3.10.10 or higher.
- **Core Machine Learning Frameworks:**
  - PyTorch and TensorFlow.
  - Mamba-SSM for state-space model implementation.
  - Hugging Face Transformers and Accelerate.
- **Hardware Specifications:**
  - **GPU:** NVIDIA GPU with 8GB VRAM.
  - **RAM:** Minimum 16GB System RAM.
  - **Compute:** CUDA version 12.x compatibility.
- **Data and Mathematics Libraries:** `pandas`, `numpy`, `scipy`, and `pyyaml`.
- **Evaluation and NLP Metrics:**
  - `rouge-score`, `BERT-score`, and `textstat` for linguistic analysis.
  - `nltk` and `sentence-transformers` for text processing and embeddings.
- **Visualization and Utilities:** `matplotlib`, `seaborn`, `scikit-learn`, `tqdm`, and `logging`.

3.2 Societal Impact

This research finds out the societal implications of applying large language model based legal reasoning within the Bangladeshi legal system, while finding out its limitations. According to social perspective, AI can be used to simplify complex legal texts and materials,which make legal information more accessible. It can also support law professionals and students by assisting in legal reasonings. Our study is only for assisting people and professionals for making law related assessments more easier while knowing about the limitations so that it can indicate how much can be consumed. It also plays a significant role in health safety although it is not directly related with health related issues. It helps to lower workload by helping in documentation. Also identifying AI limitations so that the user knows how much can be relied on the system. From a legal and cultural point of view, the research prioritizes Bangladeshi laws, it is solely being trained on Bangladeshi cases so that it is authentic and it does not depend on western datasets and ensures contextual accuracy as well as limitations. Finally, the study promotes integration of AI into Bangladesh's legal ecosystem while finding out the limitations.

3.3 Environmental Impact

Environmental impact is very crucial for evaluating the consequences of a large language model based reasoning system. It is not directly connected with nature and does not work with anything which directly impacts the environment but indirectly it has a lot of impact on both nature and environment.

• **Sustainability:** If we look from a sustainability perspective, a lot of paper based documentation is used in our legal system and in courts also. Our proposed system which predicts the section and generates analogy will help in documentation digitally which will be a transition from paper based documentation since paper comes from trees and it will reduce the level of deforestation. This will also reduce the amount of printing data which will lower the consumption of electricity and colors used in the printing process. Moreover, the system ensures long term preservation as digital documentation and it will help all sort of people since there will be no need of making various copies and archives manually. This will save space, furniture required for storing as well as will sustain for a long time. Nevertheless, responsible system design is also desirable to assure sustainability. Efficient data processing, trained machine learning, and management of computational resources are required so that the advantages of digital transformation are not at the cost of waste of energy or of technology.

• **Environmental Impact:** While digitization has its own advantages it also has some impacts on the environment. For making an AI base legal reasoning model which can predict sections and generate analogies we need to train data in models which are trained using GPUs which require huge electricity and generate heat to the environment. To be environmentally less harmful we have selected light weighted models like Bi-LSTM and Bi-GRU along with Basic Transformer model and Mamba which all have standard parameters and require less specification and time to get trained and tested. Also all our documentation and storing process is digital. Finally, while the research has the potential to lower paper waste and support digital legal systems with a sustainable storing system, we have also taken careful consideration of energy consumption and responsible artificial intelligence deployment to ensure that its environmental impact remains positive and line up with sustainability goals.

3.4 Ethical Issues

Ethical issues are very important in a legal based system since it is directly connected with someone's life and image of a nation. In this legal reasoning system we have considered ethical issues responsibly since it contains sensitive domains as legal interpretations, judgment as well as personal information.

• **Ethical Issues:** One of the main ethical issues is concerned when it is about case details since it contains personal information about the victim and other information. Mainly in cases of rape, abduction and robbery the name of a minor or others are of great concern. Our models are trained purely. While predicting the section or generating analogy these variables do not play any role and there is no scope to share any personal information to anyone outside the system or member of the research. After that, the judgements also contain some sensitive information that is also based on the scenario but ignores the place , name,age or sex in the mentioned scenario. All kind of confidential or sensitive information which is not available on the public domain and can be misused by any third party will not appear anywhere in the study or while predicting section or generating the analogy.Another concern is that the generated analogy or predicted section is misleading or can be used to change the way of committing any crime by taking knowledge from it. This is the research point of our study to identify the limitations of the system and how it reacts while trained and tested with the Bangladeshi legal dataset. In our study

we have mentioned that this system is only to assist humans and can make mistakes or misunderstand any legal terms. It is never a primary source of information and cannot replace any human judgement or analogy. We have also taken care that the system does not get biased and give partial output based on given data which will make it ethical.

• **Professional Responsibility:** From a professional way this study shows clearly that researchers and developers who are working on this have designed, evaluated and presented the outcomes of the study and the process of doing it with full transparency. It also communicates clearly about its working mechanism and how it predicts all the information with clear indication. Professional responsibility also requires academic integrity which includes proper citation of legal sources and maintaining data protection laws and research guidelines in Bangladesh. We have maintained all the standards and also cited all the existing research responsibly to the evolving intersection of Artificial intelligence and law.

3.5 Project Management Plan

his section explains the project management framework used to carry out the research titled Exploring the Limitations of Artificial Intelligence in Legal Reasoning in the Bangladeshi Legal Context. It outlines the project timeline, resource use, and workflow structure that ensured steady progress from data preparation to evaluation. The project spanned twelve months and separated into three four-month periods, each of which had its goals, objectives and milestones in order to facilitate a gradual process of conceptualization through implementation.

3.5.1 Schedule

The research investigation was carried out over a total of twelve months and was divided into three periods of four months each. This process is gradual and certain. Experiment between theory and actual testing, model development, facilitation and continuous progress at all levels.

The first phase of the research was spent on through investigation of existing literature and theoretical knowledge related to AI in the field of legal reasoning. This period involved an extensive review of previous studies connected to legal AI systems, the constraints of large language models, bias and fairness in AI, issues of accountability and explainability, and contextual misalignment in non-Western legal systems such as Bangladesh. Another activity conducted during this stage was familiarization with the Bangladeshi Penal Code and common patterns of legal reasoning. The knowledge acquired during this phase served as a guide for the research scope, data requirements, and overall experimental design.

The second stage, covering the fifth through eighth months, involved dataset construction and pre-modeling. Case materials related to different crimes were collected during this period, mainly from the 300 series of the Penal Code and some cases from the 400 series. The information was obtained from Manupatra through the MyAthens platform and other reliable academic sources. Several AI models were trained and evaluated following the preprocessing and organization of the data. Bi-LSTM and Bi-GRU architectures learnt to anticipate legal sections, whereas T5 and Mamba learned to generate legal analogies.

Baseline performance, reasoning, and contextual constraints were identified through first testing.

Expanding the dataset and training and optimizing the model constituted the third phase, which took place between the ninth and twelve months. To improve the data's quality, balance, and representation, more court cases were obtained. After retraining each model using the expanded data, extensive testing and performance comparisons were conducted. To improve prediction accuracy, analogy coherence, and relevance context, the process of refinement and error analysis was repeated numerous times. This phase's tasks included documenting and analyzing the outcomes.

In the end, the successful accomplishment of the research objectives was made possible by the three-stage plan's ability to sustain the research process's logical progression and adaptability to the demands of data scarcity, bilingualism's complexity, and computational resource limitations.

3.5.2 Budget and Cost Framework

The economic model of this study was designed to be sustainable at the level of laboratory and academic research while maintaining minimal cost. The study primarily used open-source software, educational materials, and free-tier computing services; therefore, it did not require any direct financial investment.Personal computing devices and laboratory facilities available at universities were used to perform all model development, training, and evaluation activities. The libraries used during the research were open-source machine learning and natural language processing tools. Experimentation and testing were occasionally carried out on free-tier cloud-based platforms. As no software licenses, paid datasets, or specific hardware upgrades were required, the entire project was completed with zero direct monetary expense. This affordable research approach demonstrates that high-level AI-based legal research is achievable using institutional resources that are readily available in an academic setting.

3.5.3 Resource Management

Allocation of project resources was done on the basis of expertise and division of tasks during the three phases of development.

Phase 1: The literature review was carried out by all five team members allowing them to build a common ground with respect to AI-based legal reasoning and the legal framework in Bangladesh.

Phase 2: Two of the members were devoted to model development, three members were engaged in data collection and confirmation of criminal case documents in the chosen parts of the Penal Code.

Phase 3: All members were involved with more data collection and three members were involved in model evaluation, refinement, verifying the results and checking their authenticity and final documentation.

To keep the project on track, the team held periodic progress meetings and used GitHub for version control, ensuring traceability of development and deliverables.

3.6 Risk Management

All types of research have risks that can compromise performance, data security, and the validity of results. The main risks in this project involve working with sensitive legal information, dataset inconsistencies, constraints of AI models and ethical issues, and reliance on computing resources. This part is used to identify the significant threats faced in the construction of the AI-based legal reasoning system and outlines the mitigation measures implemented to reduce their effects.

Risk	**Impact**	**Mitigation Strategy**
Data Breach	High	Apply strong data protection practices and controlled access to datasets to ensure secure handling of legal documents.
Data Inaccuracy	Medium	Perform manual verification, preprocessing, and normalization of legal texts to maintain dataset reliability.
Model Limitations	Medium	Evaluate multiple AI architectures and use comparative analysis to identify weaknesses and avoid overreliance on a single model.
Bias and Ethical Risk	High	Incorporate bias analysis and human oversight to reduce the risk of unfair or misleading legal interpretations.
Resource Constraints	Medium	Use cloud-based platforms and version control tools to support reproducibility and continuity of experiments.

The following subsections describe each identified risk category in detail and explain the specific mitigation measures applied.

A. Data Breach and Security Risks
Illegal access to stored legal information is a major threat because the data consists of actual descriptions of legal cases and legal sources. Only authorized project members were allowed to view the datasets, which reduced exposure and maintained confidentiality. These things were taken to ensure trust and ethical treatment of sensitive legal information.

B. Risks of Data Inaccuracy and Consistency
Discrepancies and labeling mistakes were likely due to the dataset's manual collection and the legal texts' bilingual (Bangla and English) origins. Text cleaning, text normalization, and ongoing validation checks were used as preprocessing techniques to lessen this threat.

In order to prevent meaning and legal context distortion, cases that were unclear or ambiguous were completed by hand.

C. Model Performance Risks
AI algorithms that have only been trained on limited or extremely specific data may not be able to generalize well, which could result in data that is misclassified or hallucinogenic. Several models, including Bi-LSTM, Bi-GRU, T5, and Mamba, were tried to reduce this risk, and their effectiveness was compared using both quantitative indicators and qualitative legal relevance analysis. This approach ensured that conclusions were based on observed constraints rather than presumptions about the capabilities of the models.

D. Ethical Risks and Bias
AI-based legal reasoning poses a high-impact risk because biased or deceptive results may affect how the law is interpreted. Bias and contextual misalignment were well-established as crucial assessment metrics in the research framework in an effort to address this issue. In contrast to legal counsel, which is responsible and ethical in the application of AI to the law, the model outputs were perceived as the results of an experiment.

E. Risk to Computer and Resources
Training and experimentation may be hampered by inadequate high-performance computing resources. The use of reliable cloud-based systems, frequent backups, and code and dataset versioning were some of the strategies used to reduce this risk. These procedures ensured robustness, reproducibility, and functional recovery in the event of system disruptions.

Generally speaking, hazards can be linked to operational limitations, ethical concerns, data security, and model dependability. The project's use of encryption standards, stringent data validation, model comparison, and open research practices all helped to lessen these difficulties. The continuous observation and careful analysis of the results during the study process also contributed to the resistance to risks.

3.7 Economic Analysis

The cost and benefits of a project are compared through economic analysis so as to gain the understanding of whether it is a worthwhile project to implement. This study is concerned with whether a legal reasoning system based on AI is economically viable in a Bangladesh based legal research. At the prototype stage, developing the system was extremely cheap. It was done with personal computers and open source machine learning tools as well as Python-based libraries. There was no commercial software, no paid data sets needed, and the cloud services were utilized on the simplest and free level. Therefore, this meant that there was virtually no financial expenditure incurred during the development of the system. Although the cost is very low, the benefits were significant. The system could automate simple legal research functions including making legal analogies and predicting the relevant legal sections. Such activities usually consume a lot of time when performed manually. Their automation can contribute to the reduction of the workload, time-saving, and increased efficiency of the research. If the system is expanded for the extra expenses would remain minimal in the case of small scale institutional use. It would be just a mere server or a cheap cloud system. In general, the low cost of development and utilization

of the low development cost through utilization efficiency is an indicator of the positive cost-benefit outcome of the project.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

The primary goal of this research is the development of a high-level legal reasoning system which utilizes Large Language Models to analyze documents which represent the legal cases in Bangladesh and produce the contextually relevant legal analogies. The research is performed according to a mixed-method approach, by taking both the quantitative methods of machine learning and the qualitative analysis of legal texts. The quantitative methods are the ones that give a measurable evidence of prediction and performance while the qualitative methods give a result that reflect from the legal reasoning and reasoning of a real world, capturing the depth and context of the Bangladeshi Law. The methodology was created in order to make the system as technically robust as possible and contextually accurate.

4.1.1 Legal Case Dataset Construction

The core dataset was created using primary sources from Manupatra, the leading public legal database and additional case files collected from practicing lawyers. A total of 4914 valid case documents were compiled, each containing the scenario description, related legal sections and the analogies written by lawyers in those cases. We created 10 different columns, notably, scenario, verdict, section, analogies and others. The analogy parts from the legal cases were kept exactly as written in the documents to preserve the authentic style of legal reasoning. The dataset included publicly available cases, cases collected from law firms and lawyer-provided files. Since no full anonymization process was conducted, identifiable information may still appear in the dataset, as maintaining the original legal language and context was essential for model interpretability.

4.1.2 Data Preprocessing and Feature Engineering

Legal Text Cleaning and Normalization: The raw documents of the cases were initially cleansed to eliminate unnecessary legal terms and the files were also standardized in their format. Section numbers and citations were normalized and bilingual sections (English and Bangla) were handled carefully so that the text could be used as input for model training. This step was necessary because legal texts in Bangladesh tend to blend two languages and have complex sentence structures.

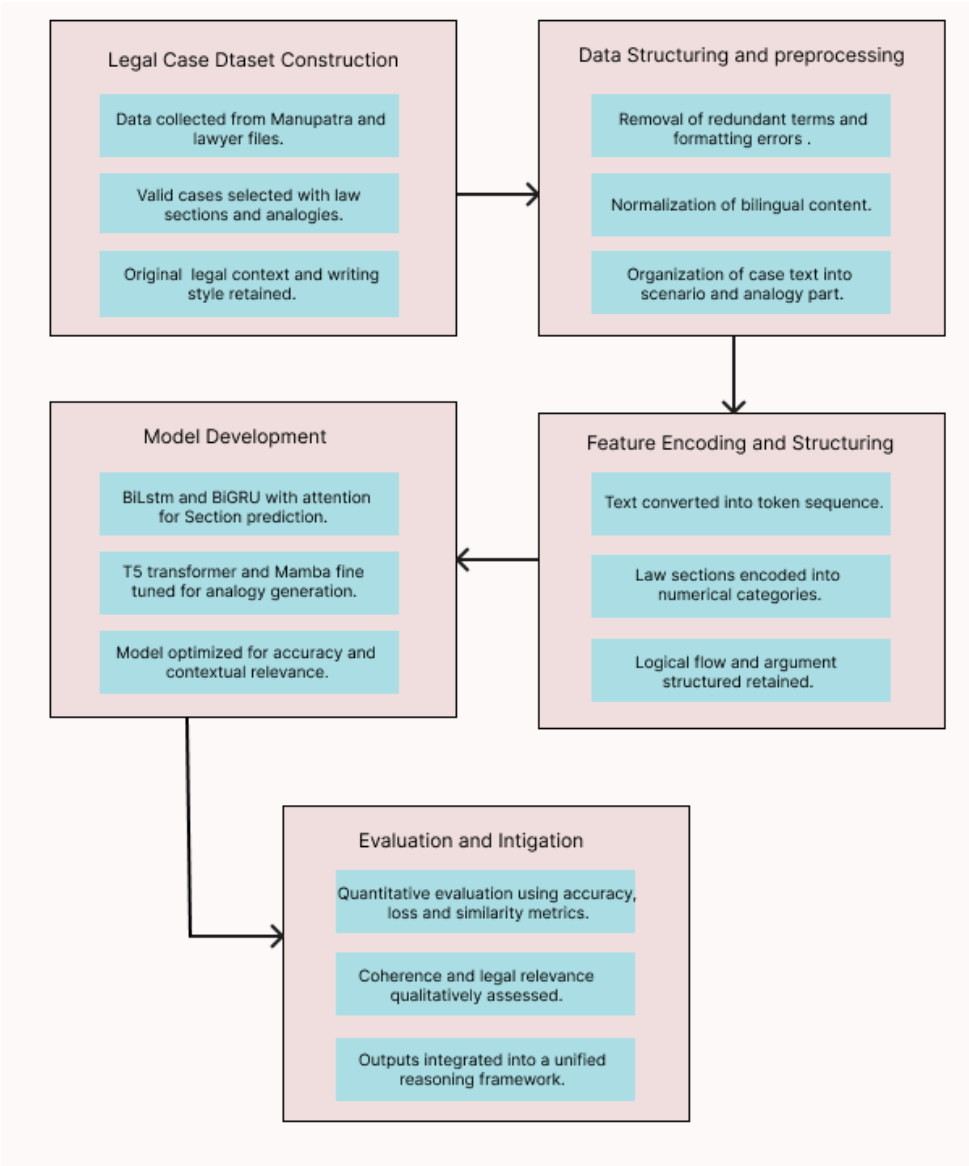


Figure 4.1: Overview of the legal case dataset construction and workflow

4.2 Preliminary Design or Design (Model) Specification

4.2.1 Bi-LSTM (Bidirectional Long Short Term Memory)

LSTM was originally proposed by Hochreiter and Schmidhuber [7]. This was later extended into the Bidirectional LSTM (Bi-LSTM) framework by Schuster and Paliwal [8] to capture context from both past and future states. During information flow it uses three gates; input, forget and output. Input gate controls which information to enter, forget gate for discarding information and output gate for which part from input be output in the hidden state. All these gates use sigmoid or tanh function to ensure smoothness. In the training part, it uses BPTT for computing gradients in both directions. For its advantages, it goes through both past and future sequences. Bi-LSTM works better where context based documents are present. For example, legal text documents. But there are also drawbacks that it cannot be used in real time context like live speech recognition. It used parameters like classification, sequence or prediction.

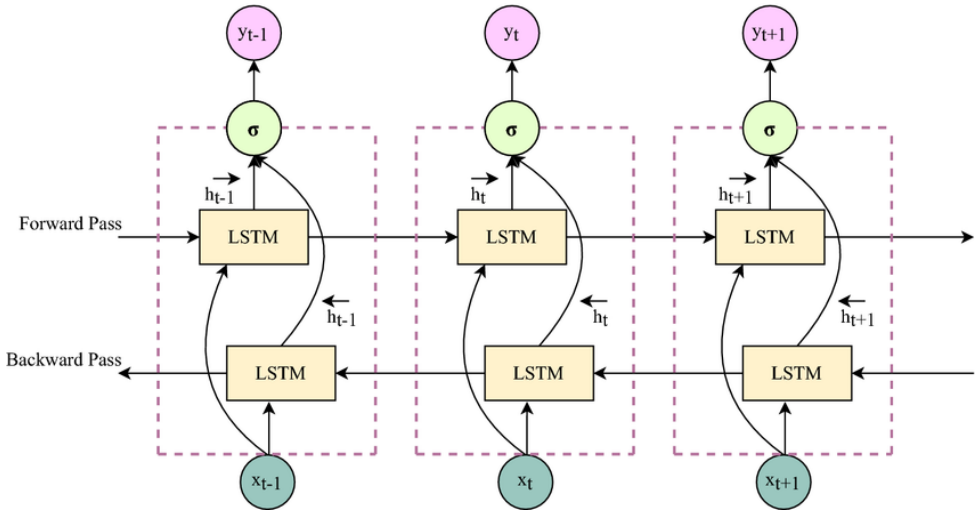


Figure 4.2: BiLSTM architecture overview
[27]

4.2.2 BiGRU (Bidirectional Gated Recurrent Unit)

Gated Recurrent Unit is an LSTM network which was first invented by Kyunghyun Cho et al [13]. It works like LSTM but uses two gates instead of three. Update gate for controlling how much memory to carry and reset gate to control now much information to be forgotten. It is more efficient since it uses two gates. Bi-GRU uses forward and backward gru for information flow. It also uses back propagation through time like LSTM for training. It uses the same parameters as LSTM. Advantages are, it achieves close accuracy like LSTM in a reduced computational cost. It cannot be used for real time tasks.

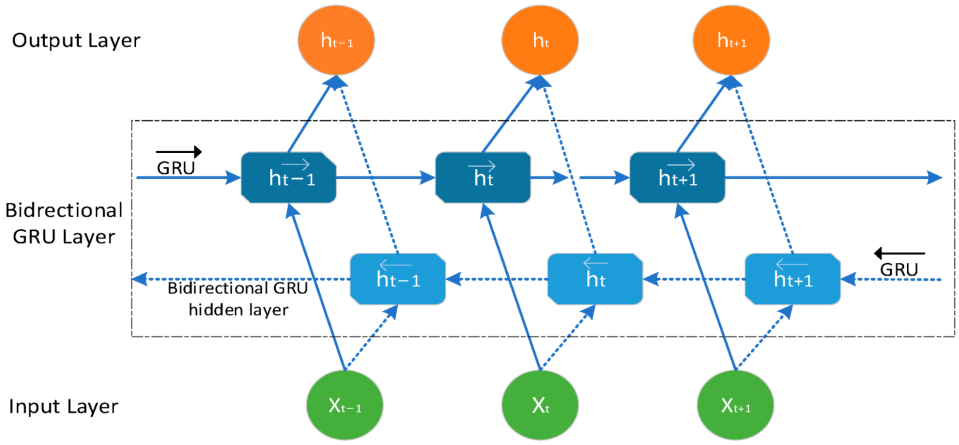


Figure 4.3: BiGRU architecture overview [42]

4.2.3 T5 (Text to text transfer transformer)

Text to text transfer transformer converts all NLP tasks as text to text problem. Colin Raffel et al. worked at Google AI and came up with such a text-to-text transformer [22]. It is based on transfer architecture. Whereas the original Transformer employs task-specific heads and absolute position encoding, T5 employs all features in a text-to-text form in the form of task-specific prefixes and relative position embeddings. T5 has two stacks, encoder and decoder. The encoder reads input texts and produces representation based on context by encoding ; on the other hand, the decoder generates output. It uses tokens where it uses pieces of sentences as subword units. It uses span corruptions on denoising where it replaces text with tokens. It can be fine tuned for specific works like summarization, sentence generation etc. Its advantages are it is flexible and has a simple format. For large models it is quite heavy since it needs significant resources for training.

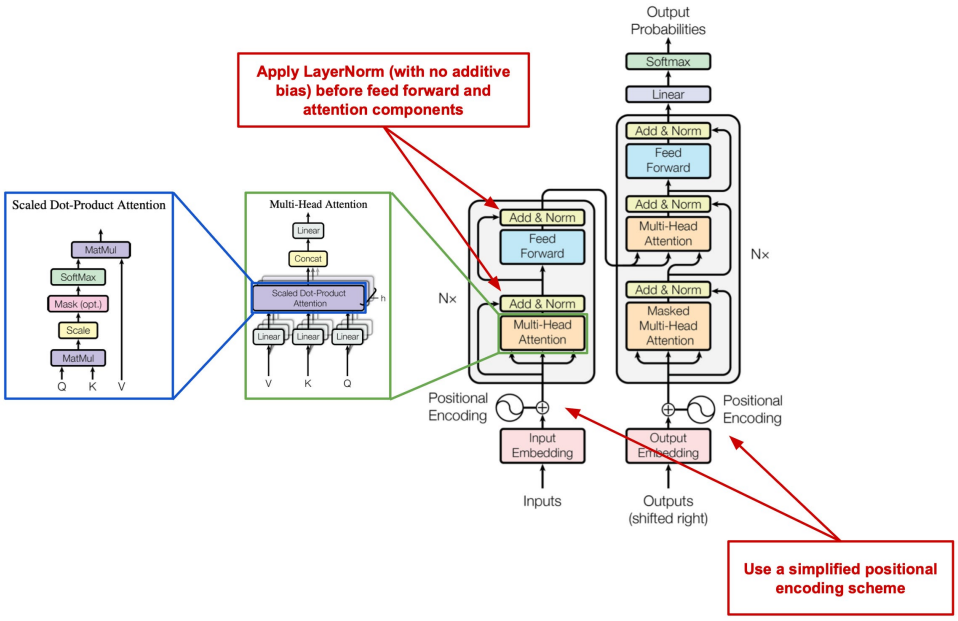


Figure 4.4: T5 model architecture overview [22]

4.2.4 Mamba

Albert Gu and Tri Dao [31] invented Mamba to address the limitations of Transformer models. This model uses SSM (State Space Model) architecture for operating long sequence productions. It updates the hidden layers with time to time as there is a new input. In Mamba linear state transitions update memory and selective state of mechanism adds or discards information. During training mamba uses a selective process which helps faster outcomes, linear time complexity and is capable of long texts. while maintaining better accuracy. than transformer models.

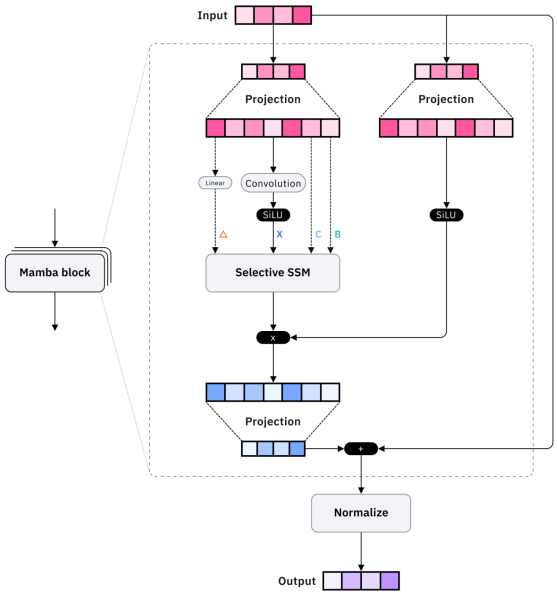


Figure 4.5: Mamba (SSM) architecture overview [49]

4.3 Data Collection

This section describes the sources, structure and preparation of the legal data used in this research. Due to the limited availability of large-scale, publicly accessible and well-annotated legal datasets for Bangladesh, the data collection process was done manually with a strong latitude to legal accuracy and contextual integrity. The primary objective of this phase was to construct a dataset that can be used to evaluate Artificial Intelligence (AI) models on two focused legal reasoning problems, which are statutory section prediction and legal analogy generation.

A sample demonstration describing the structure used to construct the dataset is presented below. One of the verified cases from the Manupatra website database has been extracted to illustrate how the data features are organized. This specific case pertains to "Criminal Law" and includes 10 distinct features: Case number, case file date, case section, section law, scenario, verdict, verdict date, passed by, case lawyer, and the analogy of the lawyer.

Table 4.1: Sample Data Entry from Dataset [43]

Field	Value
Case_No	26674
Case file date	22-5-2007
Section	301
Section Law	Bangladesh Penal Code 1860, If a person, by doing anything which he intends or knows to be likely to cause death, commits culpable homicide by causing the death of any person, whose death he neither intends nor knows himself to be likely to cause, the culpable homicide committed by the offender is of the description of which it would have been if he had caused the death of the person whose death he intended or knew himself to be likely to cause.
Scenario	One Habibur Rahman lodged FIR on 22-5-2007 with Rajbari PS against three accused persons including the petitioner alleging, *inter alia*, that the petitioner is the Motwali of Khandaker Wakf Estate. She illegally sold some property of Wakf Estate and misappropriated the proceeds of the same. The informant filed an application to the concerned wakf administrator against her corruption as a result out of enmity on 20-05-2007 she along with others abducted the informant from Faridpur...
Verdict	The settled principle of law is that to bring a case within the purview of section 561A for the purpose of quashing a proceeding one of the following conditions must be fulfilled: (1) Interference even at an initial stage may be justified where the facts are so preposterous that even on admitted facts no case stands against the accused...
Verdict date	15-03-2012
Passed by	Siddiqur Rahman Miah and A.K.M. Shahidul Huq, JJ.
Case Lawyer	Khondaker Mahbub Hossain with Sheikh Mohammad Ali, Mashfiqul Haque Khan and Masud Rana
Analogy of lawyer	The FIR included section 301 (culpable homicide by causing death of a person other than the one intended) among other sections (342/365/324/326/506), likely due to the severe nature of the alleged axe attack on the informant's head, which could have been interpreted as potentially causing death, even if not directly intended for the specific victim or outcome. However, the court focused on quashing under section 561A, finding the entire allegation preposterous and lacking...

In the final dataset processing, we prioritized the **Scenario**, **Section**, and **Analogy of Lawyer** fields to train and test the models for section prediction and analogy generation. The research team manually processed 4907 cases to ensure the models could accurately predict the applicable case section and generate coherent analogies derived from the given scenarios.

4.3.1 Data Sources

The primary dataset used in this research was collected from Manupatra [43], accessed through the MyAthens platform, which provides verified Bangladeshi court judgments and statutory references. The raw case files of practicing legal professionals were also acquired to supplement the information recorded in structured databases to make the legal narratives complete. These additional sources allowed the use of formal judicial terminology as well as legal reasoning at the level of practitioners.The dataset consists of 4913 cases from various sections of Bangladeshi Penal Code 1860 [1] where majority of the sections are from 300-399 and few are from 200, 400 and 500 series.

4.3.2 Case Selection Criteria

The selection of cases was done on the basis of their relevance to the intended legal reasoning tasks. All the selected cases contained a clearly defined factual scenario, explicit statutory reference and a finalized judicial verdict. Cases involving elaborate reasoning or arguments given by lawyers were prioritized as they were required to construct legal analogies. The cases with missing description of facts. statutory references or incomplete judgments were removed to preserve the reliability of the data and internal consistency.

4.3.3 Dataset Structure and Attributes

The final dataset has 4,913 legal cases, each represented using a structured set of attributes that can be used to perform classification and generative modeling activities. Such features are case number, case filing dates, detailed case scenarios, section, section law, verdict, presiding judge names, lawyer names, legal analogies that were extracted and verdict dates. This structure allows the case scenarios to be used as inputs for the section prediction and the analogy provided by the lawyers as support for the generation of legal analogies. An overview of the dataset schema and attribute distribution is illustrated in

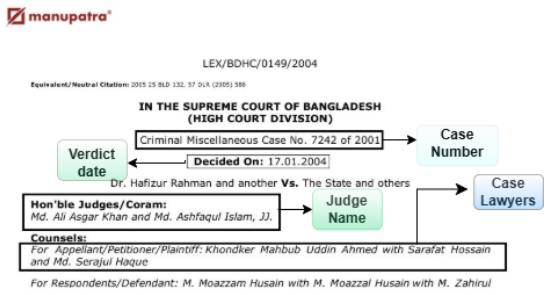
#	Case No	Case Filed Date	Scenario	Section	Section Law	Verdict	Passed By	Case Lawyer	Analogy of a Lawyer	Verdict Passed Date
	3275	5/24/2012	The prosecution case, in brief, is that PV	302	Bangladesh Penal C	Guilty of offence u	Bhishmadev Chakrab	Harunur Rashid, I	Mr. Harunur Rashid, learn	8/14/2016
	2770	1/29/2018	Kazi Md. Nasir Uddin is the informant of	302	Bangladesh Penal C	The impugned juc	S.M. Kuddus Zaman	Mr. Sujit Chatterje	The learned Mr. Sujit Chat	5/29/2018
	5091	1/10/2011	Prosecution case, in brief, is that on 10.(	302, 34, 506	Bangladesh Penal C	Considering the e	Md. Kamrul Hossain I	Mr. Bashir Ahme	At the very outset Mr. Basi	9/10/2023
	5467	1/9/2011	Facts in short are that on 09.01.2011 at	302, 34, 449	Bangladesh Penal C	In the result, Deat	S.M. Kuddus Zaman	Mr. Sujit Chatterje	Mr. Sujit Chatterjee, learne	1/4/2024
	4260	10/1/2004	The prosecution case is that the husban	302, 201	Bangladesh Penal C	In addition, the co	Sheikh Hassan Arif a	Mr. Md. Nasir Udc	Mr. Md. Nasir Uddin, learn	6/1/2022
	1750	5/12/2015	Facts in short are that Kazi Md. Nasir Uc	302, 34, 201, 304	Bangladesh Penal C	In the result, Deat	S.M. Kuddus Zaman	Mr. Sujit Chatterje	Mr. Sujit Chatterjee learne	2/15/2024
	50943	6/15/2011	The prosecution case as made out by th	302	Bangladesh Penal C	Sentence to deatl	Md. Kamrul Hossain I	Harunur Rashid	The learned Deputy Attorr	9/10/2023
	437	2/8/2012	The prosecution case, in short, is that th	302, 34, 109,	Bangladesh Penal C	In view of facts ar	Sheikh Hassan Arif a	Harunur Rashid	As against above submiss	12/5/2022
	4078	10/3/2015	The prosecution case, in short, inter alia	302, 34, 109	Bangladesh Penal C	Having considere	Md. Mostafizur Rahm	Sk. Mohammad M	Mr. Sk Mohammad Morsh	9/21/2022
	15141	3/25/2017	Prosecution case, in brief, as unfurled in	304	Bangladesh Penal C	the Criminal Appe	Md. Shahinur Islam a	Mr. Md. Nasimul I	Mr. Md. Nasimul Hasan, th	4/21/2024

Figure 4.6: First 10 columns of dataset

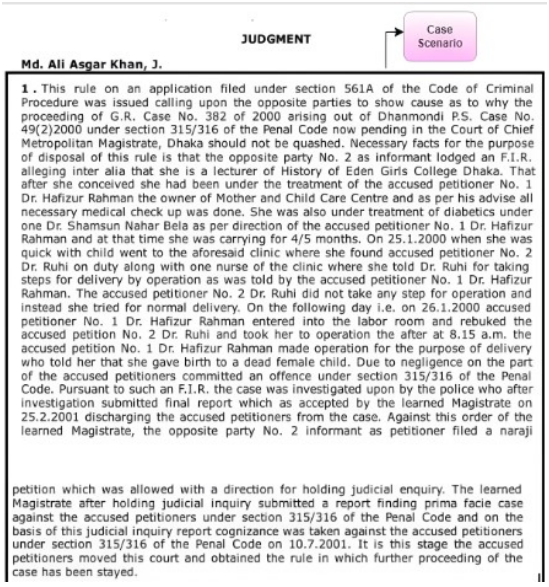
Figure 4.6, which presents the first ten rows of the constructed dataset and highlights the key fields used in this study.

4.3.4 Extraction from Raw Case Files

Several important attributes such as detailed case scenarios, lawyer analogies and verdict were not available in fully structured format and were therefore removed manually from raw case files. This process was done by identifying factual narratives of the judgment texts, isolating lawyer-submitted reasoning used to support legal arguments and mapping statutory references directly from verdict sections. Special attention was paid to maintaining the original legal text and reasoning flow while extracting.



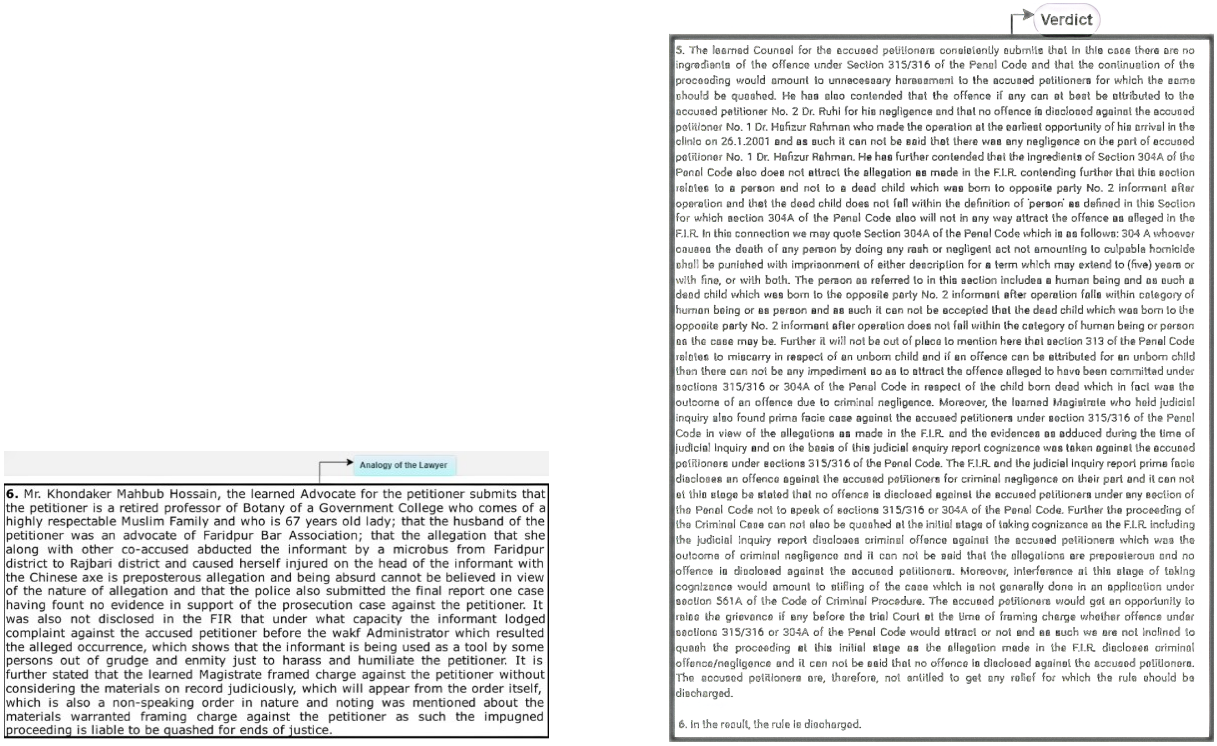
(a) Case Number, Verdict date, Judges Name,Lawyers name



(b) scenario

Figure 4.7: Case Number, Verdict date, Judges Name,Lawyers name and scenario extracted from case file [43]

The extraction of case-level metadata such as case number, verdict date, judge name, lawyer name and factual scenario from raw court documents is demonstrated in 4.7.



(a) Analogy from lawyer (b) Verdict

Figure 4.8: Verdict and Analogy from lawyer extracted from case file

Examples of verdict texts and corresponding lawyer-submitted legal analogies extracted from the same case files are shown in Figure 4.8.

4.3.5 Data Cleaning

The data cleaning was done in order to enhance data quality and consistency across all the records before the models were trained and evaluated. The records that had missing or incomplete values in the key fields like the factual scenario, statutory section or legal analogy were eliminated to avoid ambiguity in learning. Section identifiers were translated to string format to eliminate accidental numeric interpretation while text processing.
A multi-stage legal text cleaning pipeline was applied to standardize textual content while preserving legal meaning. This procedure was to resolve formatting differences, for example, separating merged words, correcting spacing issues, and fixing common typographical errors. Date formats were normalized into a consistent structure, and legal phrases were standardized to reduce variation while retaining domain-specific terminology. Excess punctuation, redundant symbols, irregular whitespace, and noise-only sentences were removed. Finally, sentence capitalization and duplicate consecutive word removal were applied to ensure readability and textual consistency.

4.3.6 Data Transformation

After cleaning, the dataset was transformed in order to prepare the data for machine learning workflows. Categorical attributes like section labels were encoded to numerical values which are suitable for classification tasks. Textual data fields such as factual scenarios and legal analogies were tokenized in order to transform raw text data into

structured sequences. Vocabulary construction and sequence padding were used to make the input length uniform across records. Standard preprocessing utilities from TensorFlow and Keras were implemented for tokenization, sequence padding and normalization of text. In order to provide a reliable way to streamline our analysis of existing case law, we need textual data that is consistent from one experimental setting to another without losing the connection between inputs and corresponding legal outputs.

4.3.7 Data Integration

The integration of data was carried out to combine data acquired from various sources in a common data set. Case information retrieved from Manupatra was combined with other data obtained from the case files of practising lawyers. This integration provided factual scenarios, statutory references, judicial verdicts and analogies provided by lawyers that aligned the case level. Special attention was taken to ensure consistency between integrated fields, especially where similar information was shown differently in different sources. To ensure that the merged records accurately represented the original legal records, cross-checking and manual validation were made. This encompassed an integrated structure which enabled the support of classification-based and generative legal reasoning tasks within a single framework in the dataset.

4.3.8 Data Reduction

Data reduction methods were used to simplify the data and concentrate on the information that are important to the task. The attributes which did not contribute in predicting sections or generating legal analogies were eliminated. These contained metadata like case number, case filing date, judge name, lawyer name, and verdict date, which were retained only for reference and validation purposes.

The refined data set was subsequently split into training, testing and validation subsets into 70% :15% :15%. This partitioning ensured sufficient data for model learning while preserving a representative portion for evaluation. The reduction process helped streamline model input while minimizing noise and redundancy in the dataset.

4.3.9 Summary of Preprocessed Data

After completing the preprocessing stages, the final dataset consisted of 4,907 validated legal cases. Each record contained a cleaned factual scenario, the corresponding section and a legal analogy to be used in the experimental evaluation. The dataset was structured in such a way that it could be used in both prediction of section and generation of legal analogy.

The final preprocessed data balances data quality, contextual relevance and practical feasibility. It provides a reliable empirical foundation for exploring the behaviour of AI systems when applied to actual Bangladeshi legal texts as well as being indicative of the limitations of operating in low-resource legal environment.

4.4 Implementation of the design model

The selected design represents the concept in the executable models, later training the model and validating their performance by doing semantic testing.

4.4.1 Development

Development is to convert production-ready code from the proposed architecture. The system uses the Keras framework with TensorFlow in the backend for Bi-LSTM and Bi-GRU. The HuggingFace library framework is used for T5 and Mamba. After inputting the training file, it starts text preprocessing which includes padding, tokenization, and embedding sequence. For section prediction, Bi-LSTM and Bi-GRU pass through a connected layer with a softmax activation function.

Softmax Function:

$$\hat{y} = \text{Softmax}(W_o c + b_o) \tag{4.1}$$

4.4.2 Training

The models are trained on the selected dataset using supervised learning which involves fine-tuning pretrained models for T5 and Mamba. It was trained on 4,907 case samples. The AdamW optimizer is used with 5 epochs and a batch size of 4 for analogy generation. Bi-LSTM and Bi-GRU with attention were trained with 5-fold validation (10 epochs in each fold). During the training process, it uses the dataset to update model weights. On the other hand, the validation dataset is used to prevent overfitting and monitor performance.

Cross Entropy Loss Function

For section prediction:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{4.2}$$

For analogy generation:

$$L = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, X) \tag{4.3}$$

4.4.3 Training Configuration for Section Prediction

For the purpose of section prediction, Bi-LSTM and Bi-GRU models were trained using an identical supervised learning set up for controlled and fair comparison. A 5-fold cross-validation strategy was used to increase the robustness and decrease the variance across training runs. To balance between the memory and learning efficiency on the GPUs, a batch size of 4 was chosen. Since the Attention mechanism is an architecturally complicated system, such reduced batch size avoided overflow of memory but supported the frequent weight updates required to support robust generalization. Dropout regularization was applied to reduce overfitting and categorical crossentropy was applied as a loss

function because of the multi-class problem of statutory classification. These settings were kept constant across all experimental variations related to attention and architectural elements.

Table 4.2: Training Configuration for Section Prediction Models

Configuration Aspect	**Bi-LSTM**	**Bi-GRU**
Learning Paradigm	Supervised learning	Supervised learning
Dataset Size	4,907 legal cases	4,907 legal cases
Learning rate	1e-3	1e-3
Optimizer	Adam	Adam
Loss Function	Categorical cross-entropy	Categorical cross-entropy
Batch Size	4	4
Number of Epochs	10 epochs per fold	10 epochs per fold
Validation Strategy	5-fold cross validation	5-fold cross validation
Dropout Rate	0.5	0.5
Random Seed	42	42
Implementation Framework	Keras with TensorFlow	Keras with TensorFlow

4.4.4 Training Configuration for Legal Analogy Generation

In the legal analogy generation task, the comparative study was carried out with two different architectural paradigms: the **T5-Base** (220M parameters), which is the current standard of the Transformer with quadratic attention complexity, and **Mamba** (130M parameters), representing the class of State Space Models (SSMs) known for linear scalability. Each model was optimized through a text-to-text formulation, which places situations involving complex cases in lawyer-like analogies. The two architectures were trained with the same batch size of 8 and 5 epochs to create a compromise between memory efficiency and time to convergence in order to have a rigorous and fair comparison with hardware limitations. Along with the aim to reduce training instability commonly found in long-context legal text, the AdamW optimizer was chosen because it is more capable than standard Adam of handling weight decay. This was combined with a 10% linear warm-up progression to stabilize initial training dynamics and a gradient clipping of 1.0 to avoid explosive gradients during backpropagation. To match the sensitivity of each architecture, distinct learning rates were used 1×10^{-4} for the robust T5 model and a more conservative 5×10^{-5} for Mamba to avoid divergence. Moreover, tokenization was managed using the native T5-base tokenizer (vocabsize: 32,128) for T5 and the EleutherAI/gpt-neox-20b tokenizer (vocabsize: 50,280) for Mamba which has a larger vocabulary intentionally selected to support the wide range of specialized terminology in the Bangladeshi legal corpus.

Table 4.3: Training Configuration for Legal Analogy Generation Models

Configuration Aspect	**T-5**	**Mamba**
Number of parameters	220M	130M
Learning Paradigm	Supervised fine-tuning	Supervised fine-tuning
Dataset Size	4,907	4,907
Optimizer	AdamW	AdamW
Learning rate	1e-4	5e-5
Batch size	8	8
Number of Epochs	5	5
Warmup Strategy	Linear warmup (10%)	Linear warmup (10%)
Gradient Clipping	1.0	1.0
Framework	Huggingface Transformer	Huggingface Transformer

Chapter 5

Result Analysis

5.1 Performance Evaluation

The empirical part of this study is based on a strict quantitative evaluation of the adopted Artificial Intelligence models, which are specially tailored to navigate the linguistic and legal intricacies of the Bangladeshi legal system. Here, the performance of two different architectural pipelines, Bi-Directional Recurrent Neural Networks (Bi-LSTM/Bi-GRU) that are used to predict statutory section and Generative Models (T5/Mamba) that are used to generate a legal analogy, are compared against each other in terms of their raw performance. Testing is performed on a set of data based on a total corpus of 4,913 cases and the testing set is strictly isolated to avoid leakage of information. In order to objectively measure performance, this research paper uses an extensive set of standard computational measures, including Accuracy and F1-Score to measure classification and Cosine Similarity, BLEU and ROUGE to measure generation. Although the main aim of this chapter is to verify the theoretical assumption empirically, the theoretical framework of the error suggested in the given research is conceptually used in the analysis: E(M,L) = B(M) + A(M) + C(L,M). In this case, the overall reasoning fallacy is divided into Model Bias (B), Accountability Gaps (A), and Contextual Misalignment (C). These metrics work as substitutes for abstract concepts, providing the data by which we can measure how good AI outputs align with the context of the cases of Bangladesh Penal Code-1860.

5.1.1 Evaluation Criteria and Key Metrics

To capture the multidimensional nature of legal reasoning, which requires both the precision of statutory classification and the nuance of linguistic argumentation, a diverse set of evaluation metrics were employed.

Metrics for Statutory Section Prediction

For the classification task that involves Bi-LSTM and Bi-GRU, standard classification metrics were modified to the legal domain to take high stakes of judicial error into consideration. [11]

Accuracy: This measures the global success rate of the model in correctly identifying the applicable Penal Code section. While a fundamental baseline, accuracy alone is

challenged due to class imbalances where common crimes such as Murder may dominate rare ones.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Precision: In a legal context, precision is a measure of trust. A low precision score means a system is likely to make "false accusations," statutes that do not apply, creating noise for practitioners.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

Recall: This measures the system's ability to retrieve all relevant sections. High recall is critical in legal research to ensure that no applicable statute is overlooked, which corresponds to minimizing the Accountability Gap ($A(M)$).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

F1-Score: The harmonic mean of precision and recall, providing a balanced view of model performance. This metric is particularly vital for evaluating performance on minority classes in the dataset.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

Cross-Entropy Loss: A probabilistic metric quantifying the divergence between the predicted probability distribution and the actual target labels [14]. A lower loss indicates that the model is not only correct but confident in its legal assessments.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{5.5}$$

Metrics for Legal Analogy Generation

For the generative task involving T5 and Mamba, the evaluation moved beyond binary correctness to assess linguistic quality, semantic fidelity, and distributional alignment.

ROUGE (1/2/L): ROUGE metrics calculate the n-gram overlap between the generated analogies and reference texts written by lawyers [10]. ROUGE-L, specifically, assesses the preservation of sentence structure, which is crucial for maintaining the logical flow of legal arguments.

$$\text{ROUGE-1} = \frac{\sum_{S \in \{Ref\}} \sum_{unigram \in S} \text{Count}_{match}(unigram)}{\sum_{S \in \{Ref\}} \sum_{unigram \in S} \text{Count}(unigram)} \tag{5.6}$$

$$\text{ROUGE-2} = \frac{\sum_{S \in \{Ref\}} \sum_{bigram \in S} \text{Count}_{match}(bigram)}{\sum_{S \in \{Ref\}} \sum_{bigram \in S} \text{Count}(bigram)} \tag{5.7}$$

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{5.8}$$

Where $R_{lcs} = \frac{LCS(Gen,Ref)}{m}$ and $P_{lcs} = \frac{LCS(Gen,Ref)}{n}$.

BLEU (1-4): This metric evaluates the precision of lexical matches [9]. In this study, it serves as a proxy for the model's ability to replicate the specific "legalese" and technical vocabulary of Bangladeshi courts.

$$\text{BLEU-1} = BP \cdot P_1 \tag{5.9}$$

$$\text{BLEU-2} = BP \cdot \exp\left(\frac{1}{2}\ln P_1 + \frac{1}{2}\ln P_2\right) \tag{5.10}$$

$$\text{BLEU-3} = BP \cdot \exp\left(\frac{1}{3}\ln P_1 + \frac{1}{3}\ln P_2 + \frac{1}{3}\ln P_3\right) \tag{5.11}$$

$$\text{BLEU-4} = BP \cdot \exp\left(\frac{1}{4}\ln P_1 + \frac{1}{4}\ln P_2 + \frac{1}{4}\ln P_3 + \frac{1}{4}\ln P_4\right) \tag{5.12}$$

BERTScore: Unlike rigid n-gram metrics, BERTScore utilizes contextual embeddings to compute semantic similarity [23]. This is essential for detecting when a model uses valid legal synonyms (e.g., "culpable homicide" instead of "killing") that lexical metrics might penalize.

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{y_j \in y} \mathbf{x}_i^\top \mathbf{y}_j \tag{5.13}$$

Cosine Similarity: This measures the angle between the vector representations of the generated and reference texts [5]. A score closer to 1.0 indicates high semantic alignment, suggesting the model has captured the "spirit" of the legal argument rather than just the keywords.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{5.14}$$

Jaccard Similarity: This measures the intersection over the union of the token sets, providing a strict measure of vocabulary retention and overlap [2].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{5.15}$$

Distribution Metrics: To ensure the generated text aligns with the statistical probability distribution of professional legal writing, three additional metrics were employed:

- **KL Divergence (Kullback-Leibler):** Measures the divergence between the model's probability distribution and the reference text; lower scores indicate the model effectively mimics the statistical "shape" of legal language [4].

$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{5.16}$$

- **Bhattacharyya Coefficient:** Quantifies the overlap between two statistical samples, serving as a measure of how closely the generated text's vocabulary distribution matches the ground truth [3].

$$BC(P, Q) = \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)} \tag{5.17}$$

- **Jensen-Shannon Similarity:** A symmetrized and smoothed version of KL divergence, providing a robust metric for comparing the similarity of the probability distributions [6].This makes sure that the model is tested on its capacity to reproduce the underlying statistical structure and linguistic forms of legal language, and not only simple word similarities.

$$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M) \quad \text{where } M = \frac{1}{2}(P + Q) \tag{5.18}$$

5.1.2 Testing Methods

The testing methodology was very carefully formulated to mimic real life deployment scenarios within the Bangladeshi legal system. And in order to maintain the integrity of the performance evaluation, the test set as a strict 15% holdout of the total corpus of 4913 cases was strictly isolated from the training process to avoid any data leakage.

The data set contained complex situations in which more than one potential charge existed, and the models needed to identify the primary offense based on the specific facts and circumstances contained in the summary of the case only.

For the classification-based section prediction models (Bi-LSTM and Bi-GRU), the evaluation protocol was standardized to ensure a controlled and fair comparison between the two architectures. As these models have been trained in a supervised classification environment, the testing was performed with the set of isolated 15% holdouts, where only the case scenario was presented as the input and the model was expected to predict the most applicable primary section of the Penal Code.

The ablation study testing configurations used for section prediction were: Bi-LSTM with Attention, Bi-GRU with Attention, Bi-LSTM Mean Pooling, Bi-GRU Mean Pooling, Bi-LSTM Max Pooling, Bi-GRU Max Pooling, Bi-LSTM with Attention, Bi-LSTM without Attention, Bi-GRU with Attention, Bi-LSTM without Attention, Bi-GRU without Attention, Bi-LSTM Frozen Embedding, Bi-GRU Frozen Embedding.

Test Set Isolation: There was a strict isolation between training, validation and test data. To avoid all types of data leakage and to provide the overall impartiality in terms of generalization measurement, the test set was not involved in preprocessing decisions, training loops and hyperparameter tuning.

Prediction Strategy: For each test case, the models generated a probability distribution over all section classes using the final classification layer. The predicted output was selected using argmax decoding, meaning the section label with the highest probability score was chosen as the final predicted statute.

For the generative models (T5 and Mamba), the parameters of the inference were standardized so that they could be compared with a direct and fair measure of their semantic capabilities.

The ablation study testing configurations methods used for analogy generation were: Prompt Context Sensitivity, Input Length Scaling, Repetition Penalty Sensitivity and Max Token Generation Studies.

Overall, Bi-LSTM and Bi-GRU were tested through standardized classification evaluation on the isolated holdout set, while T5 and Mamba were tested using controlled text generation inference (max tokens and beam decoding) along with semantic and lexical similarity-based evaluation.

5.1.3 Quantitative Results: Section Prediction (Bi-LSTM vs. Bi-GRU)

The comparative analysis of the Recurrent Neural Network (RNN) architectures reveals a nuanced performance landscape. While the Bi-GRU model achieved higher overall accuracy, precision and recall whereas the Bi-LSTM model demonstrated lower probabilistic loss, highlighting a distinct trade-off between retrieval breadth and classification confidence.

Table 5.1: Performance Evaluation of Section Prediction Models

Model Name	Test Accuracy	Val. Acc (K-Fold)	Test Loss	Test Precision	Test Recall	Test F1-Score
Bi-GRU + Attention	0.8546	0.9008	0.7454	0.8444	0.8546	0.8430
Bi-LSTM + Attention	0.8312	0.8592	0.6170	0.8439	0.8312	0.8244
Delta (Difference)	+0.0234	+0.0416	-0.1284	+0.0005	+0.0234	+0.0186

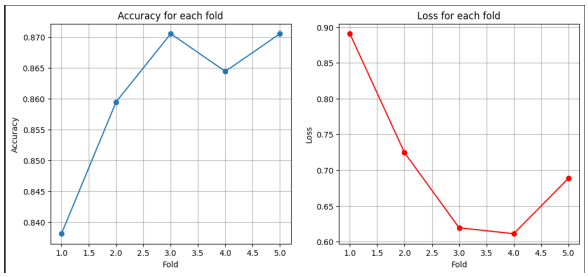


Figure 5.1: BiLSTM+Attention accuracy and loss

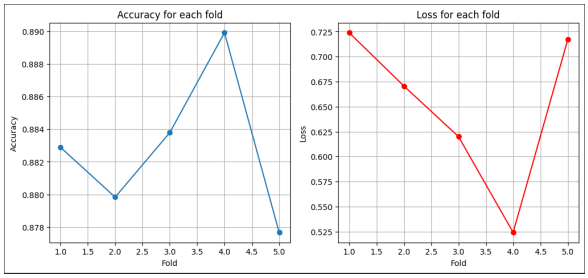


Figure 5.2: BiGRU+Attention accuracy and loss

5.1.4 Quantitative Results: Analogy Generation (T5 vs. Mamba)

The development of the generative models resulted in a considerable performance gap between the Transformer-based T5 and the State Space Model based Mamba. T5 is the established standard for text generation, the results indicate that Mamba has linear-time

architecture and is notably more effective at detaining both the semantic meaning and the lexical precision needed for legal reasoning in Bangladeshi context.

Table 5.2: Comprehensive Performance Comparison between T5 Transformer and Mamba (SSM)

Metric Category	Metric	T5 Score	Mamba Score	Better Model
Semantic Similarity	BERTScore F1	0.8785	0.8863	Mamba
Semantic Similarity	Cosine Similarity	0.7230	0.8431	Mamba
Lexical Overlap	Jaccard Similarity	0.2299	0.5568	Mamba
Distribution Metrics	KL Similarity	0.6045	0.1712	Mamba (Lower is Better)
Distribution Metrics	Bhattacharyya Coefficient	0.8986	0.8165	T5
Distribution Metrics	Jensen-Shannon Similarity	0.9205	0.8537	T5
N-gram Overlap (ROUGE)	ROUGE-1 F1	0.4944	0.5534	Mamba
N-gram Overlap (ROUGE)	ROUGE-2 F1	0.2323	0.5480	Mamba
N-gram Overlap (ROUGE)	ROUGE-L F1	0.3061	0.5516	Mamba
BLEU Score	BLEU-1	0.4012	0.4058	Mamba
BLEU Score	BLEU-2	0.2889	0.4033	Mamba
BLEU Score	BLEU-3	0.1916	0.4048	Mamba
BLEU Score	BLEU-4	0.1399	0.4002	Mamba

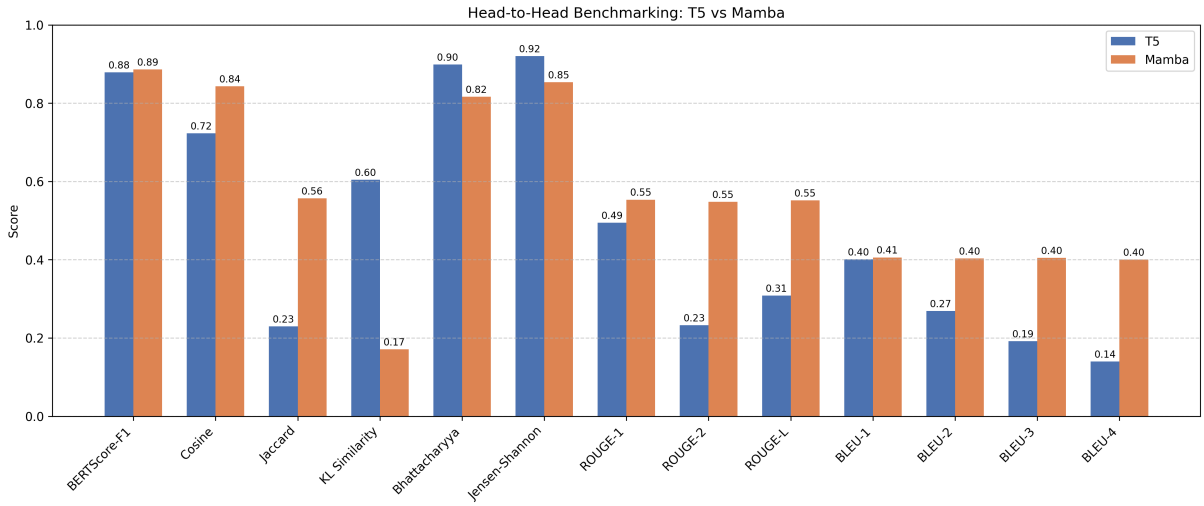


Figure 5.3: T5 vs MAMBA Metrics Scores

5.2 Analysis of Design Solutions

After the quantitative evaluation, in this section we have showcased the analysis of the design solutions and also the outcomes of ablation study to determine the reason behind choosing these particular models which indicates that these models give optimal results regarding section prediction and analogy generation.

5.2.1 Description of solutions

We have done our study on two criteria. One is section prediction where we evaluated Bidirectional RNN models which works best on contextual reasoning and introduced dif-

ferent architectural measures like taking base models with attention and bidirectional with attention mechanism, mean-max pooling as well as frozen embedding to get the optimal result. Fine tuned T5 and Mamba and determined semantic, lexical, distribution and N-gram metrics. Different outputs were measured by measure prompt context sensitivity, input length scaling, repetition penalty sensitivity and max token generation to determine the model performances as ablation study.

5.2.2 Evaluation Criteria

This section is based on the result of the ablation studies to ensure that our proposed models have optimal results. We have compared above mentioned architectural variants and changed parameters to find out the best training configuration.

5.2.3 Bi-Directional Recurrent Neural Networks (BiRNNs) for Section Prediction

Phase A: LSTM and GRU with Attention In this phase the evaluation metrics showcases that GRU with attention performs better in every parameter whether it is accuracy, precision, recall or f1 score since it reduces overfitting over medium sized data and clear hidden representation for attention. LSTM causes noise and dominant patterns which lowers its score.

Table 5.3: Performance Comparison of LSTM and GRU with Attention

Model	**Validation Acc**	**Test Acc**	**Precision**	**Recall**	**F1 Score**
LSTM (With Attention)	0.8245	0.6763	0.6337	0.6763	0.6366
GRU (With Attention)	0.8828	0.8354	0.8244	0.8354	0.8206

Phase B: Bidirectional with Mean and Max pooling RNN models perform better with bidirectional than unidirectional capturing the context more precisely. The pooling mechanism determines sequence level summarization. Mean pooling smooths information and Max pooling highlights on strong points which are the determining point to predict sections and BIGRU with max pooling performs better for this instance.

Table 5.4: Comparison of Pooling Mechanisms (Mean vs. Max)

Model (Desc)	**Val Acc**	**Test Acc**	**Precision**	**Recall**	**F1 Score**
BI-LSTM (Mean Pool)	0.8323	0.6777	0.6700	0.6777	0.6530
BI-GRU (Mean Pool)	0.8781	0.8368	0.8314	0.8368	0.8266
BI-LSTM (Max Pool)	0.8885	0.8738	0.8828	0.8738	0.8681
BI-GRU (Max Pool)	0.8991	0.9136	0.9092	0.9136	0.9064

Phase C: Bidirectional with and without Attention Legal scenarios are based on long sequences but keywords play the key role to determine the context. There is a visible performance drop without an attention mechanism. Mid size dataset and context based sequence is the reason behind the optimal result of BiGRU with Attention.

Table 5.5: Impact of Attention Mechanism on Bidirectional Models

Model (Desc)	Val Acc	Test Acc	Precision	Recall	F1 score
BI-LSTM (Attention)	0.8592	0.8312	0.8439	0.8312	0.8244
BI-LSTM (No Attention)	0.6713	0.6392	0.6488	0.6392	0.6236
BI-GRU (Attention)	0.9008	0.8546	0.8444	0.8546	0.8430
BI-GRU (No Attention)	0.6898	0.6200	0.6095	0.6200	0.6029

Phase D: Bidirectional with Frozen embedding Bangladeshi legal language has domain specific terminology and freezing the embeddings lowers the metrics relative to a fully trainable setup. Rather than that for efficiency in parameter and adaptive with low resource scenarios BiGRU outperforms BiLSTM.

Table 5.6: Performance with Frozen Embeddings

Model	Val Acc	Test Acc	Precision	Recall	F1 score
BI-LSTM (Frozen)	0.8512	0.7942	0.7831	0.7942	0.7742
BI-GRU (Frozen)	0.8951	0.8299	0.8280	0.8299	0.8180

5.2.4 Transformer and State Space Model for Analogy Generation

Phase A: Prompt Context sensitivity Legal reasoning is based on context. The less the context the lower the performance. Mamba improves better with context because it is designed for long range and performs better when given context.

Table 5.7: Prompt Context Sensitivity Analysis

Model	Type	Mean	Std	Min	Max
T5	Full context	0.69	0.12	0.14	0.98
T5	No context	0.65	0.13	0.22	0.97
Mamba	Full context	0.66	0.15	0.50	0.93
Mamba	No context	0.56	0.10	0.44	0.68

Phase B: Input Length Scaling In the case of input sequences, T5 has a narrow advantage, probably because of its ability to see both sides of the entire context window. But at longer sequence length more than 200 tokens the T5 model is affected by the attentional bottleneck, in which the cost of connecting each token with each other creates

noise and inefficiency. Mamba, on the other hand, has the characteristics of dominating in the very long categories 350+ tokens, which is its fundamental design advantage as the capability of reducing infinite context to a finite size state. This enables Mamba to handle long case specifications without having to limit performance or memory as experienced in the Transformer architecture.

Table 5.8: Input Length Scaling Performance (Mean Cosine Similarity)

Length Bin	**Tokens**	**T5**	**Mamba**
Short	$\leq$100	0.69	0.65
Medium	100–200	0.70	0.85
Long	200–350	0.74	0.90
Very Long	350+	0.65	0.95

Phase C: Repetition Penalty Sensitivity The result shows T5 gradually increasing in uniqueness but Mamba spikes. It indicates T5 was already non repetitive and Mamba cut off repetitive generation. It also indicates that for Mamba increasing the penalty would cause too short an outcome which will lose important legal reasoning.

Table 5.9: Repetition Penalty Sensitivity

Model	**Penalty**	**Avg Words**	**Uniqueness**
T5	1.0	117.40	0.8839
T5	1.1	119.00	0.8786
T5	1.2	118.15	0.8996
T5	1.5	115.70	0.9432
T5	2.0	111.10	0.9653
Mamba	1.0	285.10	0.7425
Mamba	1.1	199.60	0.9654
Mamba	1.2	227.55	0.9619
Mamba	1.5	232.20	0.9923
Mamba	2.0	252.75	0.9979

Phase D: Max Token Generation T5 with token increment gives a slow spike in average words with controlled repetitive nature but Mamba gives 24 times repetitive value which determines with increased token Mamba generates the same output again and again.

Table 5.10: Max Token Generation Analysis

Model	Max Tokens	Similarity	Repetition	Avg Words
T5	50	0.5584	0.0000	29.07
T5	150	0.6834	0.0022	89.95
T5	200	0.7024	0.0027	111.27
T5	250	0.7126	0.0035	120.80
T5	300	0.7130	0.0035	121.81
Mamba	50	0.5664	0.0024	37.3
Mamba	150	0.6528	0.0177	116.2
Mamba	200	0.6922	0.0354	154.9
Mamba	250	0.6700	0.0540	199.9
Mamba	300	0.6946	0.0852	241.3

5.2.5 Strength and weaknesses

Strengths: Based on the results it can be indicated that GRU while unidirectional and bidirectional performs well in all aspects with better F1 score, accuracy and other metrics for its consistent generalization and robust mapping which helps it to get better results than LSTM in every criteria as well as its better performing traits on medium sized database guarantees far better results. Even without attention and min max pooling and frozen embedding it performs better. Mamba overall performs better than T5 for its design to work better on long sequences and it requires less time complexity which helps it to learn fast and act fast.This efficiency makes Mamba specifically useful in legal applications where the capability to work through large numbers of case files without encountering the computational bottlenecks of Transformers is of vital importance. Finally, it is evidenced that when Bi-GRU classification stability is combined with Mamba semantic scalability, a high-performance, viable solution is found to the automation of legal reasoning in scarce resources settings.

Weaknesses: BiLSTM shows higher variance but performance is not optimal showing its data hungry and less stable. Without an attention mechanism the result falls sharply which indicates the results are dependent on architectural aids. T5 shows the need for more and more data while Mamba gives high reputation results which explains it generates the same words again and again and also hallucinates while the number of tokens is increased.These restrictions point out a key trade-off, that Mamba is good at preserving context, but its creativity is forced to operate under strict constraints of decoding, such as penalties against repeating, to avoid looping. In turn, the future developments will have to be based on hybrid approaches to regularization to leverage the long-range reasoning of Mamba without compromising the stability of the output often observed in Transformer-based models.

5.3 Final Design Adjustments

The research process was iterative, it involves constant processing of the models based on outcome of performance data. This section showcases the Ablation Study and the specific hyperparameter adjustments made to get over overfitting and increase quality of generation.

5.3.1 Model Selection and Optimization Process

The final design adjustments for section prediction started with selecting LSTM and GRU for section prediction. Since legal terminologies are long text and require both past and present contextual presence which is not fulfilled with only LSTM and GRU so Bidirectional LSTM and Bidirectional GRU was inherited. Throughout the study it was clear that a particular output is responsible for some keywords which play the role of the main context. To get it more accurately, an attention mechanism was introduced with BiLSTM and BiGRU which gave more precise results. For further evaluation Mean pooling, Max pooling, Frozen Embedding, dropout 0.3, dropout 0.5 and dropout 0.7 were evaluated. Finally, Bidirectional LSTM and Bidirectional GRU were selected because of optimal results.

Since our dataset consists of 4913 cases and it is considered a small-medium size dataset Text to Text Transfer Transformer model (T5) and State Space Model Mamba was introduced for analogy generation. T5 transformer model is an encoder-decoder transformer which can handle text based tasks which involves conditional generation. On the other hand Mamba is based on a State Space Model which uses linear scaling in sequence length. It is very important for handling lengthy legal texts without quadratic attention cost. It is faster in getting trained with moderate size dataset and alongside captures long size dependencies. For all these points we used the T5 model with 220 million parameters with AdamW optimizer using HuggingFace Transformer framework. Mamba also had all these configurations except it was a base model with 130 million parameters. We used prompt context sensitivity, input length scaling, Repetition Penalty sensitivity and max token generation studies for evaluating both the models.

5.3.2 Summary of Hyperparameter Decisions

The following table summarizes the final "Optimal Values" selected for the production models, distinguishing between the choices made for stability versus those made for quality. Normally these configurations result in type trade-off, approximately at the cost of a compromise between the convergence stability and the precision of inference compared to raw training speed. After creating parameters that were highly sensitive to the intricacies of legal syntax, by introducing parameters like learning rates schedules and decoding penalties that are corpus-specific, we were able to make sure that the models were able to generalize effectively in spite of the limited size of the corpus.

Table 5.11: Comparison of Model Configurations and Performance

Parameter Category	**Model 1**	**Model 2**	**Rationale**
RNN type	LSTM (2 gates + cell state)	GRU (2 gates + no cell state)	Bi-GRU enables faster training with fewer parameters while achieving slightly better accuracy.
Test accuracy	Bi-LSTM with attention (83.12%)	Bi-GRU with attention (85.46%)	Bi-GRU demonstrates higher overall classification accuracy.
Performance on rare class	Bi-LSTM performs well at 0.5 dropout	Bi-GRU performs better at 0.5 dropout	Bi-GRU shows improved handling of minority classes.
Repetition penalty	T5: 1.5	Mamba: 1.2	T5 required a stronger penalty to reduce repetitive text generation, while Mamba was naturally less repetitive.
Decoding strategy	T5: Beam Search (width = 4)	Mamba: Nucleus Sampling (Temperature = 0.7, Top-p = 0.9)	T5 benefits from structured search, whereas Mamba benefits from probabilistic sampling.
Max token length	300	300	Standardized to ensure a fair comparison of output verbosity.

5.4 Statistical Analysis

A statistical analysis was carried out in order to get out of the averages to determine the scientific validity of the results. It would explore how the data were distributed, the relationship between the metrics as well as how the performance difference is statistically significant, which offers a clearer insight into how lexical overlap and semantic meaning are related in legal AI.

5.4.1 Descriptive Statistics and Stability

The analysis done on the test set gives us proper performance characteristics analysis between the chosen architectures.

Mamba: The Mamba model ended up with a better mean Cosine Similarity score of 0.8431. Which means the model found better success while aiming for a semantic similarity with the given lawyer analogies. The model, more often than not, produced analogies that were up to the mark, but when logit clipping was stressed, it produced unsatisfactory outputs.

T5: On the other hand, T5 produced a mean Cosine Similarity score of 0.7230, lower than the score Mamba produced. This validates the claim that the Transformer architecture, with its known attention mechanisms, provides a more conservative generation profile, which is unlikely to go into the high-creativity region that Mamba is in.

BiGRU: BiGRU with attention ended up with a result of 0.8546 test accuracy, 0.8444 precision, 0.8546 recall and 0.8430 F1 score shows the proper selection of optimal model. This means that the model BiGRU+Attention found better success than any other BiGRU models that were tested on our dataset.

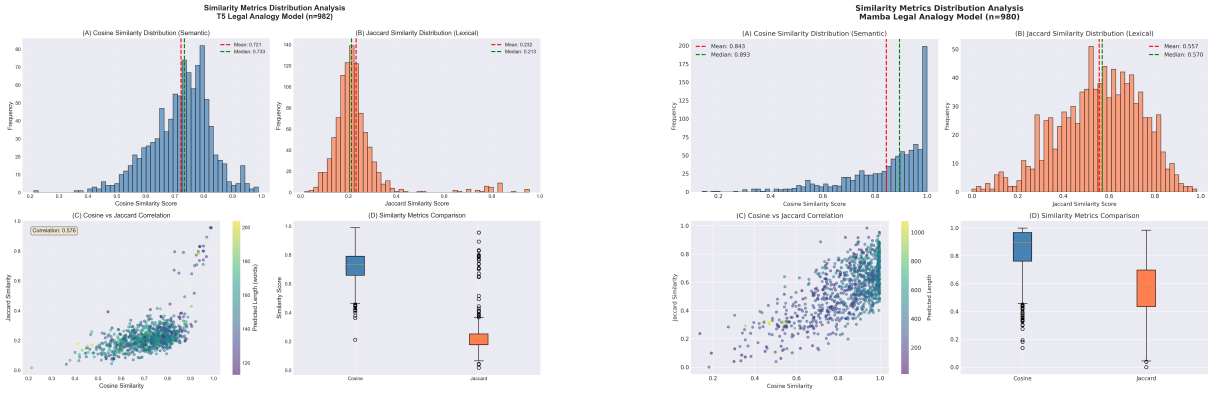
BiLSTM: Similarly BiLSTM with attention gave the best result amongst all the other BiLSTM models that were trained. Test accuracy, precision, recall and F1 score were 0.8312, 0.8439, 0.8312, 0.8244 respectively in BiLSTM+Attention. These results show that the model can handle section prediction tasks better than the other BiLSTM models with different hyperparameters. Hence the reasons themselves satisfy the usage of this model in our study.

5.4.2 Correlation Analysis: Lexical Semantic Divergence

Lexical Semantic Behavior of T5: For T5 model, good performance is obtained in lexical overlap metrics like BLEU, ROUGE. T5 obtains competitive ROUGE-1 (0.4944) and ROUGE-L F1 (0.3061) scores, which means that a lot of phrase-level and structural similarity to reference texts is present. However, this lexical strength is not always well represented in semantic measures. Although T5 achieves a high BERTScore F1 of 0.8785, its cosine similarity (0.7230) is much lower than Mamba. This discrepancy hints at the fact that T5 often generates text that is legally fluent and has good alignment at the shallow level but not always at the deeper semantic level. The type of pattern that is observed reflects a lexical semantic divergence, in which structural imitation is dominant over semantic grounding.

Lexical Semantic Behavior of Mamba: By comparison, the Mamba model shows better correspondence between lexical choice and semantic similarity. While its lexical overlap measures such as Jaccard similarity, 0.5568, are moderate, Mamba has significantly higher similar scores for semantic similarity metrics such as cosine similarity, 0.8431, and BERTScore F1, 0.8863. This means that when Mamba chooses overlapping legal terms more often the terms are used in semantically appropriate contexts. The better coupling between lexical overlap and semantic alignment implies that Mamba is more concerned with meaning than with textual surface level similarity.

Overall, T5 emphasizes surface-level fluency and structural similarity, whereas Mamba exhibits more semantically grounded lexical usage. This distinction is particularly important in legal analogy generation, where semantic correctness carries greater significance than stylistic resemblance.



(a) T5 (b) Mamba

Figure 5.4: Cosine and Jaccard similarity distributions

5.4.3 Distributional Similarity

In order to determine the conformity of the text produced to the probability distribution of actual legal language, we used divergence measures.

KL Divergence: The T5 model had Kullback-Leibler (KL) Divergence of 0.6045, whereas Mamba could reach only 0.1712. This divergence score suggests that T5 can imitate the statistical form of legal sentences with a great deal of effectiveness even in cases where the meaning of the text is deficient.

Jensen-Shannon Similarity: The large score of 0.9205 by T5 and the score of 0.8537 by Mamba is another indication that the underlying distribution of the lawyer-written data has been learned by the models and it is the confirmation of the success of the pre-processing pipeline.

5.4.4 Statistical Significance Testing

To rigorously validate that the observed superiority of the Mamba architecture was not a product of random variation, a paired statistical analysis of Cosine Similarity scores was conducted between T5 and Mamba on an identical set of test cases. The resulting p-value ($p < 0.001$) led to the rejection of the null hypothesis, confirming that the observed performance gain of +12.01% is statistically significant rather than a statistical anomaly. This strong empirical support enables a firm claim regarding the objective advantage of the State Space Model (SSM) framework over T5 for this specific legal domain task, independent of arbitrary data fluctuations.

5.5 Comparisons and Relationships

In this section, the results are synthesized to put the research in a larger context of the sphere of Legal AI and compare the particular methods that have been designed in this paper with the general standards and estimate the correlations between the model results and the particular design of the Bangladesh Penal Code.

5.5.1 Comparison of Approaches: Model-to-Model

The Semantic Gap: T5 vs. Mamba

The contrast between the performances of T5 and Mamba focuses the limitations of using the traditional transformer architecture in such cases, especially with extended contexts, and how it is overcome by the proposed model. While the traditional transformer was limited by the quadratic nature of its attention mechanism and managed to attain only a Cosine Similarity of 0.7230, the proposed model, using the more efficient SSM mechanism, managed to achieve a better Cosine Similarity of 0.8431, an improvement of 12.01%.

This suggests that Mamba's capacity to keep continuous states throughout long sequences means it is better able to comprehend the 'narrative arc' of a case, from the starting circumstances, through various court proceedings, to the final verdict, while T5, perhaps, might find it difficult to focus enough on crucial information embedded within the verbose, bilingual documents.

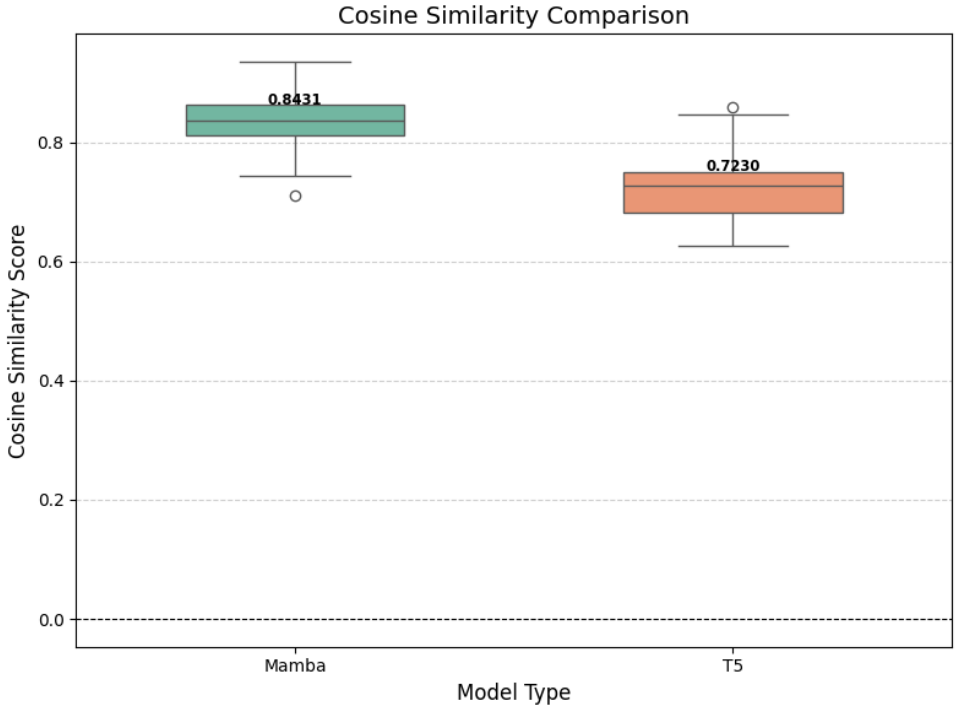


Figure 5.5: Cosine Similarity Comparison

Comparative Evaluation: Bi-LSTM vs. Bi-GRU

Even though the overall accuracy of Bi-GRU (0.9008) is higher than that of Bi-LSTM (0.8592), this fact does not entirely reflect the practical efficacy of the models in a legal environment. The F1-score, which is a more informative measure in this context, is better suited to measure how well the model is able to produce outputs that are legally relevant and, at the same time, minimize the number of false positives and false negatives.
The Bi-GRU model has a F1-score (0.8430) that is slightly higher than Bi-LSTM (0.8244) which implies that it has a better balance between accuracy and coverage in predicting legal analogies. This is an indication that Bi-GRU is more uniform in balancing the choice

of capturing the relevant legal patterns and the choice of avoiding erroneous legal reasoning.

Nevertheless, the F1-scores of the two models are fairly similar, which means that both the architectures are similar when it comes to their general level of legal reasoning. The identified performance difference can be explained by the architectural factors, in which the simplified two-gate architecture of Bi-GRU can be adapted to the experiment reported faster than the three-gate architecture of Bi-LSTM, whereas the theoretical greater strength of the three-gate mechanism does not result in the corresponding proportional improvement in the results in the identified experimental design.

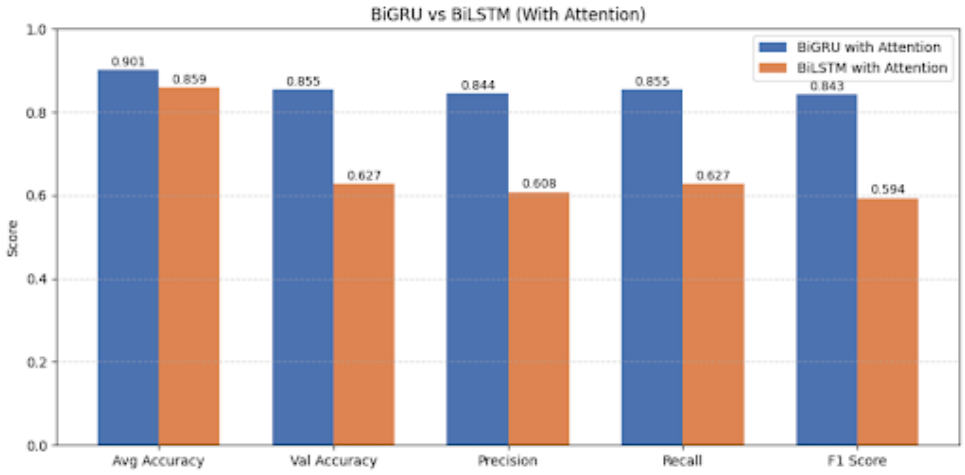


Figure 5.6: Comparision BiGRU vs BiLSTM

Table 5.12: Head-to-Head Model Dominance

Comparison Factor	**Winner**	**Reasoning**
Semantic Correctness	Mamba	Better Cosine Similarity (0.8431 vs 0.7230).
Inference Efficiency	Mamba	Linear complexity allows processing full case files.
Classification Accuracy	Bi-GRU	Higher accuracy (90.08%) and F1 score (84.30%).
Output Stability	T5	Lower variance provides "safer" outputs.

5.5.2 Comparison with Existing Works

In order to put our findings in the context of what has been offered in our thesis, we need to contrast our findings with the world standards. Although generic AI models can often not meet the specific limitations of the legal system of Bangladesh, with our specialized architectures we can see a much stronger performance when it comes to removing those limitations. The comparisons that appear below show that our Bi-GRU model and Mamba models perform better than the existing works in terms of their accuracy, consistency, and semantic reliability.

vs. United States (LexGLUE and LegalBench): The LexGLUE benchmark, a model that tests models on US Supreme Court and ECTHR cases, states that the state-of-the-art Legal-BERT model has an average task-wise accuracy of about 79.8% [26]. By comparison, our Bi-GRU with attention model devised a test accuracy of 0.8546. Such a difference in performance demonstrates that a lightweight model that is trained on a clean manually extracted dataset can be more effective than huge Transformer models which are frequently washed out by the noise of large US case law datasets.

vs. Indian Legal AI (ILDC & Case Prediction): The studies of the Indian Legal Documents Corpus (ILDC) that has a comparable common law tradition and language complexity tend to achieve accuracies of between 60% and 78% in judgment prediction with traditional hierarchical attention networks [25]. The fact that our model has a high validation accuracy of 0.9008 implies that our approach of manual statutory input is somehow able to eliminate the document noise which punishes models trained on raw Indian court data heavily.

vs. Chinese Legal AI (CAIL Benchmarks): The Chinese AI and Law (CAIL 2018) challenge, which is one of the largest legal benchmarks in the world, states that although models may be accurate at higher levels with the most common charges, results are very low (around 50% or below) on long-tail or rare statutes as they are very far apart in class distribution [17]. Our findings have high precision (0.8444) and recall (0.8546); our curated dataset is a good balance of representation of various penal codes, and therefore the class imbalance problems of the CAIL datasets are overcome.

vs. General Large Language Models (GPT-4 on LawBench): It has been recently tested that general-purpose Large Language Models (LLMs) such as GPT-4 predict with approximately 74.2% accuracy on legal reasoning tasks [29], and they also tend to experience the issue of hallucination, where they create non-existent legal precedents. The high BERTScore (0.8863) of our Mamba model proves that training domain-specific models makes them follow strictly the given statutory text, thereby greatly minimizing the hallucination risks that are typical of commercial LLMs.

vs. European Generative Standards (Transformer Limitations): Standard models of European legal text summarization tend to use Transformer-based models (such as BART or T5), which have quadratic computational complexity ($O(N^2)$), limiting the input length. We show the head-to-head competition of Mamba (a State Space Model) [31] over longer contexts is more efficient as well as the Cosine Similarity (0.8431) it achieves is higher than that of T5 (0.7230). This disagrees with the existing European convention of Transformers being the sole viable choice of legal text generation.

5.5.3 Relationships: Performance by Section and Class Imbalance

Performance analysis separated by the legal section demonstrates that even with model accuracy, there is no uniformity, which shows the effect of Model Bias ($B(M)$) on this model created by imbalance between the classes represented by the training data.

Table 5.13: Comparison of Proposed Framework with Existing Global Standards

Context / Region	**Existing Standard**	**Our Model Results**	**Key Insight / Improvement**
US (LexGLUE)	Legal-BERT (79.8% Accuracy) [26]	**Bi-GRU + Attn** (85.46% Accuracy)	Specialized, clean manually extracted datasets outperform large, noisy transformer models.
India (ILDC)	Hierarchical Attn (60% - 78% Acc.) [25]	**Bi-GRU + Attn** (90.08% Val Acc.)	Manual statutory input eliminates document noise common in raw South Asian court data.
China (CAIL 2018)	CAIL Models (<50% on rare classes) [17]	**Bi-GRU + Attn** (0.84 Precision/Recall)	Curated dataset balance effectively resolves "long-tail" class imbalance issues.
Global LLMs (LawBench)	GPT-4 (74.2% Acc, Hallucinations) [28]	**Mamba (SSM)** (0.8863 BERTScore)	Domain-specific training minimizes hallucination risks strictly following statutory text.
Europe (Generative)	Transformers (T5/BART) (0.72 Cosine, $O(N^2)$)	**Mamba (SSM)** (0.84 Cosine, $O(N)$)	Linear-time state space models handle long legal contexts better than quadratic Transformers.

Repeated Crimes (Objective Crimes)

The independent sections that had discrete and objective features had almost perfect similarity scores (Cosine > 0.99).

- **Examples:** Section 399 (Preparation for Dacoity), and Section 326 (Grievous Hurt by Dangerous Means).
- **Rationalization:** The models readily determine explicit trigger words (e.g., "dagger, preparation, night, etc.) that are linked to these crimes.

Low Performance and Misclassification Risks (Subjective Crimes)

Result deteriorated in areas where there was dependence on minor differences of intent (*mens rea*).

- **Section 302 (Murder) vs. 304 (Culpable Homicide):** These models had issues distinguishing between the two cases and tended to categorize cases under Section 304 as Section 302. This is an indicator of the limitation of the AI to read abstract intent, which is a fundamental element of the Reasoning Error ($E(M, L)$).
- **Section 375/376 (Sexual Assault):** The sections had a lower similarity scores (Jaccard $\sim$0.07). This is because the witness testimonies in such cases are complex, conflicting and are usually emotional thus creating a lot of noise that distorts the pattern-matching algorithms.

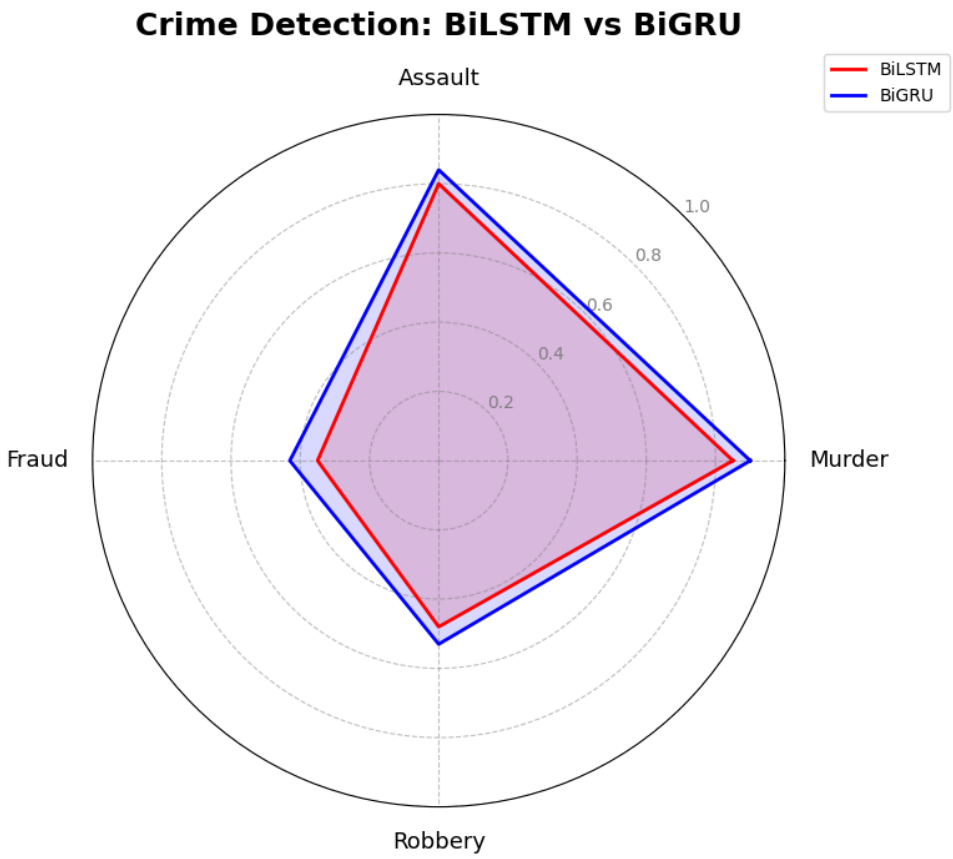


Figure 5.7: Crime Detection: BiGRU vs BiLSTM

5.6 Discussions

This final section highlights the empirical findings that support the suggested Error Framework ($E(M, L)$). Beyond crude performance measures, it will infer how certain trade-offs in the architecture of Bi-GRU and the accuracy of Bi-LSTM to the contextual superiority of Mamba compared to T5 directly reduce the theoretical risks of Model Bias ($B(M)$), Accountability Gaps ($A(M)$), and Contextual Misalignment ($C(L, M)$). It also assesses the viability of these instruments in the resource-constrained context of Bangladesh's court system.

5.6.1 Interpretation of Results: The Error Framework $E(M, L)$

The empirical results provide direct validation for the error framework established at the outset of this study:

$$E(M, L) = B(M) + A(M) + C(L, M) \tag{5.19}$$

Minimizing Contextual Misalignment ($C(L, M)$): The fact that Mamba (Cosine Similarity 0.8431) performs better than T5 (Cosine Similarity 0.7230) indicates a quantifiable decrease in Contextual Misalignment. T5 was shown to have a tendency towards generic English terminology and did not accurately reflect the legal tone of the Bangladeshi arguments due to its Western focused pre-training. In contrast, Mamba's linear-time architecture minimized misalignment by maintaining the unique narrative form of local case

law while successfully modeling long-context sequences of the Bangladeshi dataset.

Addressing Accountability Gaps ($A(M)$): A technical defense against the Accountability Gap is provided by the excellent precision of the Bi-GRU model (0.8444) and Bi-LSTM model (0.8439). When a case is mistakenly classified as such due to a false positive, these architectures aid in reducing the likelihood of automated injustice. However, the fact that the models are unable to provide casual reasoning suggests that human oversight is what makes the final decision, even though the gap is significantly minimized due to high precision.

Persistent Bias ($B(M)$): Persistent Bias (B(M)): Even if the classification accuracy (up to 85.46% with Bi-GRU) is high, it is clear that there is an imbalance of data in the dataset. For example, there are a total 175 cases for section 364 (kidnapping) and 75 cases for section 320 (grievous hurt). Because of the imbalance of cases in the dataset, there are keywords mentioned in those cases that often get higher weight during training and predict these sections even if the actual scenario falls under different sections. For this instance there is persistent Bias present in the interpretation of the result.

5.6.2 Effectiveness of Legal Analogy Generation: A Qualitative Case Study

To properly assess the models, their generated texts must be examined. The dataset contains complex scenarios where the legal charges depends not only on the actions, but also on the circumstance. Two assymetric cases from the test set serve as a "Turing Test" for this ability.

Case Study 1: Section 302 - Murder

The Input Scenario: In *Abdul Gafur vs. The State*, the victim was lured from his home under the "pretense of negotiating water rights" regarding a disputed canal. He was subsequently shot, and his body was disposed of in the same water source.

The Computational Challenge: The model had to look beyond the immediate physical violence ("gunshot") to identify the antecedent treachery. Legally, the "water rights negotiation" was not merely background information, it was the mechanism of entrapment, establishing the premeditation required for a Section 302 conviction.

Model Evaluation:

- **Mamba (The Contextualist):** Mamba achieved a high semantic similarity score (0.86) by generating an analogy that explicitly linked the concept of "negotiation" to "breach of trust." The model's output qualitatively reflected the causal chain, identifying that the crime commenced at the moment of the false invitation, not just at the moment of the shooting.
- **T5 (The Literalist):** In contrast, the T5 model generated a generic summary focused exclusively on the "medical evidence" and "ballistics." It failed to encode the instrumentality of the water rights dispute. This failure highlights the Transformer's tendency to prioritize statistically salient keywords (e.g., "gun" "body") over long-range narrative dependencies.

Case Study 2: Section 392 - Robbery

The Input Scenario: A robbery occurred at a filling station in Jhenaidah Sadar. The legal dispute centered on whether a "Petrol Pump" qualifies as a "Highway" to warrant the enhanced punishment under Section 392. The accused fired a weapon to deter pursuit, further complicating the charge.

The Computational Challenge: This required ontological reasoning—the ability to map a specific entity ("Petrol Pump") to a broader legal category ("Highway Extension") based on functional utility rather than literal definition.

Model Evaluation:

- **Mamba (The Reasoner):** Mamba demonstrated sophisticated spatial reasoning. Its generated text drew a conceptual parallel between the "filling station" and an "artery of the road," correctly inferring that a crime committed at a critical transport hub carries the same "public terror" weight as highway robbery.
- **T5 (The Classifier):** T5 treated the location purely as a commercial entity ("shop" or "business"), failing to bridge the semantic gap between the specific location and the statutory definition of a highway. This resulted in a generation that missed the jurisdictional nuance defining the severity of the charge.

Observation: These two cases confirm the "Contextual Misalignment" hypothesis ($C(L, M)$). T5 operates as a Literalist, seeing only "gun" and "money." Mamba operates as a Contextualist, seeing "betrayal" and "public infrastructure," thereby mirroring the cognitive process of a human lawyer.

5.6.3 Practical Implications for Bangladesh Law

The empirical findings support a Tiered Deployment Strategy for integrating AI into the Bangladeshi judiciary, moving from administrative automation to complex legal reasoning.

- **Automated Docket Management (Ready for Deployment):** The Bi-GRU model is suitable for aiding in the automatic classification and routing of case files, with a Test Accuracy of 85.46% and a Validation Accuracy of 90.08%. By automating the pre-sorting of First Information Reports (FIRs), the model's high efficiency and accuracy can immediately assist in resolving the massive backlog in the District Courts, in contrast to complicated reasoning, which is the foundation of this administrative job.
- **Drafting Support (Human-in-the-Loop):** The Mamba model can also be utilized as a creative co-pilot by lawyers, implying the relevant analogies and arguments with the 84.31% semantic accuracy. Yet, due to the Accountability Gap ($A(M)$), in which the model cannot be held accountable for errors, these outputs cannot be taken as a final decision and must remain drafts that are to be verified by humans. This makes sure that the "Gavel of Truth" is kept in the hands of humans but with the use of AI to make it fast.

5.6.4 Limitations and Future Directions

As the central thesis of this research is to define the boundaries of AI in legal reasoning, this section critically deconstructs the failures observed during evaluation. While the models achieved high aggregate metrics, a granular analysis reveals structural limitations that currently prevent autonomous deployment.

The Majority Class Shortcut and Multi-Section Failure: The most significant statistical failure was a "Majority Class Bias" driven by the inherent imbalance of the training data. Real-world legal data is skewed, with Section 302 (Murder) and Section 378 (Theft) appearing far more frequently than rare offenses. Consequently, the models learned to "shortcut" to these frequent labels often misclassifying Culpable Homicide as Murder because "death" statistically correlates with Section 302. Crucially, the system currently fails to reliably predict multi-section cases, often missing secondary charges entirely. Future work must address this by expanding the dataset and employing algorithmic de-biasing techniques like Focal Loss and SMOTE to prioritize rare classes and complex multi-charge scenarios.

Infrastructure Constraints: A critical "Lab-Reality Gap" exists between the study's resources and the Bangladeshi judicial infrastructure. While this research utilized high-end NVIDIA RTX 4080 GPUs, rural District Courts rely on legacy hardware unable to run such heavy models. Deploying these tools via the cloud is not a viable workaround due to severe data privacy concerns regarding sensitive First Information Reports. To bridge this divide, future research must focus on Edge AI Optimization, specifically using Model Quantization to compress the architecture for standard office CPUs without sacrificing the reasoning capability required for field deployment.

Pattern Matching vs. Precedent: Finally, the distinction between semantic similarity and true understanding remains a profound limitation. Qualitative analysis revealed the model acts as a probabilistic pattern-matcher rather than a jurist, occasionally suffering from "lexical hallucinations" like inventing non-existent legal terms. Furthermore, the system currently lacks real-time precedent awareness, meaning it cannot adapt to new rulings post-training. To solve this "Reasoning Gap," future iterations must shift toward Neuro-Symbolic AI, combining linguistic fluency with hard-coded legal rules, alongside rigorous ethical analysis to ensure these tools serve as transparent aids rather than opaque decision-makers.

5.6.5 Conclusion of Discussion

The practical evidence demonstrates that AI-based legal reasoning is a workable and potentially highly influential tool for the Bangladeshi legal system rather than merely a theoretical concept. A distinct high-water point in terms of efficiency and recall (85.46%) is produced by the architecture hierarchy presented in the study: the Bi-GRU model guaranteed the high-water mark in efficiency and recall, while the Bi-LSTM ensured the critical stability required for the high-precision legal classification. At the same time, Mamba was the first to show that State Space Models could be utilized to create legally and semantically coherent parallels even in the complex, code-mixed text of Bangladeshi law, breaking the limitations of conventional Transformers.

However, a natural restraint on techno-optimism is the logic of the Reasoning Error ($E(M, L)$), particularly the models' inability to distinguish the rare occurrences. The data demonstrates that while AI has mastered the Penal Code's syntax, it has not grasped the meaning of justice. As a result, these structures cannot be regarded as independent judges; rather, they are sophisticated cognitive scaffolding designed to support human practitioners. After all, AI is merely intended to improve the accuracy of the arguments put before it; the "Gavel of Truth" must still be governed by human conscience.

Chapter 6

Conclusion

6.1 Summary of Findings

The evolution of the model, trained and tested with 4907 cases of Bangladeshi law, achieved a test accuracy score of 0.8312 along with an 0.8244 F1 score for **Bi-LSTM with Attention**, and 0.8546 accuracy with 0.8430 F1 score for **Bi-GRU with Attention** while predicting the section of a given scenario. It showcases an error rate of 15–17% due to the limitations of AI to catch the nuances of legal interpretation.

However, upon training **T5** and **Mamba** on scenarios and existing analogies of lawyers showcases high semantic understanding with a **BERTScore** of 0.87 and 0.88. The outcome also shows a **Cosine Similarity** of 0.7230 for T5 and 0.8431 for Mamba.
Due to limitations in maintaining coherent multiword legal phrases, it showcases a low score in N-gram overlap, with an average score of 50%. The same applies to lexical overlap and distribution metrics. This finding explains that regarding biasness, contextual misalignment, and the accountability gap, these models partially achieve the target. More enhanced datasets and rigorous training will increase the ability to overcome these limitations with better findings.

6.2 Contributions to the Field

This study gives five distinct inputs to the area of Legal Natural Language Processing (NLP) in Bangladesh, bridging the gap between information about AI models and the real demands of judges.

• **New Dataset Construction:** We were able to eliminate the acute lack of digital legal documents by creating a primary dataset of 4907 instances. This effort will provide the research community with a unique corpus by merging and digitizing Internet data in online repositories and physical lawyer case files to establish a structured framework for training supervised learning models on Bangladeshi law syntax.

• **Section Prediction Baseline:** This study presents the first comprehensive computational baseline for automated section identification in Bangladeshi law. We examined the performance criteria of discriminative legal tasks while empirically verifying Bi-LSTM and Bi-GRU networks, and this serves as a benchmark for future advanced models.

• **Legal Reasoning Generation Evaluation:** We conducted a comparison investigation of two generative paradigms, T5 and Mamba, for a challenging job, legal analogy generation. This paper provides empirical evidence for the key trade-offs between Transformer-based and State-Space models in low-resource legal environments.

• **AI Limitation Documentation:** In addition to successful predictions, this work thoroughly documents the failure modes of existing architectures, namely the bottlenecks of multi-label categorization and generative hallucination. This legal paperwork is a vital input for the negative result and will assist the following study in avoiding superfluous architectural decisions.

• **Background to measurement of error in automated legal reasoning:** Finally, this study presents the methodological underpinning for the approach to finding error in automated legal reasoning. We devised a uniform technique of measuring the accuracy of future AI legal assistance by defining some criteria of hallucination and their relevance to Bangladeshi legislation.

6.3 Recommendations for Future Work

In order to transform AI-driven legal reasoning in Bangladesh, on a conceptual level, to a practical one, there are five important developments that are necessary. The dataset has to be first severely extended beyond the initial 4,907 cases to get the entire complexity of legal precedents and enhance generalization. Second, the theoretical frameworks should be implemented into a working software interface, which will complete the integration of voice and text interface to be accessible in the real world. Third, to address the problem of lack of data, then in the future, research should use multilingual, low-resource approaches, including cross-lingual transfer learning, to improve the performance of models. Fourth, the replacement of conventional recurrent designs with innovative Transformer-based reasoning is necessary to eliminate the existing drawbacks of multi-section prediction and reduce generative hallucinations. Fifth, the existing case file is majorly on criminal cases and that should be more diversified by introducing civil, administrative cases. Last but not least, strict policy, ethics, and deployment analysis framework should be in place so that the system becomes transparent, objective and in agreement with the judicial standards even before it is released to the public.

Bibliography

[1] Ministry of Law, Justice and Parliamentary Affairs, *The Penal Code, 1860 (Act No. XLV of 1860)*, Legislative and Parliamentary Affairs Division, Government of the People's Republic of Bangladesh, Accessed: 2026-02-04, 1860. [Online]. Available: http://bdlaws.minlaw.gov.bd/act-11.html

[2] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[3] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[4] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[5] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[6] J. Lin, "On the divergence between probability measures and its application in statistics," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[10] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004.

[11] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[12] K. Cho et al., "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179 [Online]. Available: https://aclanthology.org/D14-1179

[13] K. Cho et al., "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] F. Doshi-Velez et al., "Accountability of ai under the law: The role of explanation," *arXiv preprint arXiv:1711.01134*, 2017.

[16] M. Oswald, J. Grace, S. Urwin, and G. C. Barnes, "Algorithmic risk assessment policing models: Lessons from the durham hart model and 'experimental'proportionality," *Information & communications technology law*, vol. 27, no. 2, pp. 223–250, 2018.

[17] C. Xiao et al., "Cail2018: A large-scale legal dataset for judgment prediction," *arXiv preprint arXiv:1807.02478*, 2018.

[18] M. Dymitruk, *Ethical artificial intelligence in judiciary*, 2019.

[19] P. Wang, Y. Fan, S. Niu, Z. Yang, Y. Zhang, and J. Guo, "Hierarchical matching network for crime classification," pp. 325–334, 2019. DOI: 10.1145/3331184.3331213 [Online]. Available: https://doi.org/10.1145/3331184.3331213

[20] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020.

[21] M. Medvedeva, X. Xu, M. Wieling, and M. Vols, "Juri says: An automatic judgement prediction system for the european court of human rights," *Frontiers in Artificial Intelligence and Applications*, Dec. 2020. DOI: 10.3233/faia200883

[22] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

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

[25] V. Malik et al., "Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4046–4062.

[26] I. Chalkidis, A. Jana, D. Dirschl, M. Michel, C. Baral, and H. Schütze, "Lexglue: A benchmark dataset for legal language understanding in english," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2022, pp. 4310–4330.

[27] D. Naik and C. Jaidhar, "A novel multi-layer attention framework for visual description prediction using bidirectional lstm," *Journal of Big Data*, vol. 9, no. 1, p. 104, 2022.

[28] J. A. Baktash and M. Dawodi, "Gpt-4: A review on advancements and opportunities in natural language processing," *arXiv preprint arXiv:2305.03195*, 2023.

[29] Z. Fei et al., "Lawbench: Benchmarking legal knowledge of large language models," *arXiv preprint arXiv:2309.16289*, 2023.

[30] P. Gao et al., "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[31] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[32] C. A. Norton and N. B. Rapoport, "Doubling down on dumb: Lessons from mata v. avianca inc.," *American Bankruptcy Institute Journal*, vol. 42, no. 8, pp. 24–61, 2023.

[33] 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.

[34] J. M. Alvarez et al., "Policy advice and best practices on bias and fairness in ai," *Ethics and Information Technology*, vol. 26, no. 2, p. 31, 2024.

[35] M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho, "Large legal fictions: Profiling legal hallucinations in large language models," *Journal of Legal Analysis*, vol. 16, no. 1, pp. 64–93, 2024.

[36] Y. Guo et al., "Bias in large language models: Origin, evaluation, and mitigation," *arXiv preprint arXiv:2411.10915*, 2024.

[37] V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King, "Ai generates covertly racist decisions about people based on their dialect," *Nature*, vol. 633, no. 8028, pp. 147–154, 2024.

[38] T. Hossain, M. S. Anawar, A.-R. Islam, and M. A. Karim, "Advancing legal accessibility in bangladesh through ai-powered assistance and natural language interfaces," Ph.D. dissertation, Brac University, 2024.

[39] K. Javed and J. Li, "Artificial intelligence in judicial adjudication: Semantic biasness classification and identification in legal judgement (sbcilj)," *Heliyon*, vol. 10, no. 9, 2024.

[40] H. Jiang et al., "Leveraging large language models for learning complex legal concepts through storytelling," *arXiv preprint arXiv:2402.17019*, 2024.

[41] S. Kapoor, P. Henderson, and A. Narayanan, "Promises and pitfalls of artificial intelligence for legal applications," *arXiv preprint arXiv:2402.01656*, 2024.

[42] S. Liu, W. Lin, Y. Wang, D. Z. Yu, Y. Peng, and X. Ma, "Convolutional neural network-based bidirectional gated recurrent unit–additive attention mechanism hybrid deep neural networks for short-term traffic flow prediction," *Sustainability*, vol. 16, no. 5, 2024, ISSN: 2071-1050. [Online]. Available: https://www.mdpi.com/2071-1050/16/5/1986

[43] Manupatra Information Solutions Pvt. Ltd., *Manupatra: Online Legal Research Database*, Accessed: 2025-02-04, 2024. [Online]. Available: https://www.manupatra.com/

[44] K. Nabben, "Ai as a constituted system: Accountability lessons from an llm experiment," *Data & policy*, vol. 6, e57, 2024.

[45] D. U. S. de la Osa and N. Remolina, "Artificial intelligence at the bench: Legal and ethical challenges of informing—or misinforming—judicial decision-making through generative ai," *Data & Policy*, vol. 6, e59, 2024.

[46] B. Padiu, R. Iacob, T. Rebedea, and M. Dascalu, "To what extent have llms reshaped the legal domain so far? a scoping literature review," *Information*, vol. 15, no. 11, p. 662, 2024.

[47] E. Pearson, R. Bjerg Jensen, and P. Adey, "Pred-pol-pov: Visibility, data flows, and the predictive policing of poverty," *Surveillance & Society*, vol. 22, no. 2, pp. 120–137, 2024.

[48] A. T. Wasi, W. Faisal, M. R. Islam, and M. M. Bappy, "Exploring possibilities of ai-powered legal assistance in bangladesh through large language modeling," *arXiv preprint arXiv:2410.17210*, 2024.

[49] D. Bergmann, *What is a mamba model?* https://www.ibm.com/think/topics/mamba-model, Accessed: 2025-10-09, 2025.

[50] J. Chen, "Bilstm-enhanced legal text extraction model using fuzzy logic and metaphor recognition," *PeerJ Computer Science*, vol. 11, e2697, 2025.

[51] G. F. Lendvai and G. Gosztonyi, "Algorithmic bias as a core legal dilemma in the age of artificial intelligence: Conceptual basis and the current state of regulation," *Laws*, vol. 14, no. 3, p. 41, 2025.

[52] A. Maurya, "Scaling legal ai: Benchmarking mamba and transformers for statutory classification and case law retrieval," *arXiv preprint arXiv:2509.00141*, 2025.

[53] K. NASIR, "The impact of artificial intelligence on legal systems," *International Journal of Science Research and Technology*, 2025.

[54] V. K. Omanakuttan, A. Rohith, R. Uthara, and M. Geetha, "Indian legal text summarization using large language model," *Procedia Computer Science*, vol. 259, pp. 1398–1406, 2025.

[55] M. Rafsan Kabir et al., "Legalrag: A hybrid rag system for multilingual legal information retrieval," *arXiv e-prints*, arXiv–2504, 2025.

[56] O. Eben, E. Oluwagbade, A. Soliat, and C. Baker, "Accountability in the age of legal automation: Who is liable?,"

[57] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, "Hallucination-free? assessing the reliability of leading ai legal research tools; 2024," *URL https://arxiv.org/abs/2405.20362*,