INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

LLM-Backed Chatbot Lawyer for Enhanced Legal Services in Sri Lanka

A Project Specification & Design Prototype Document by

Omar Salman Shiraz

Supervised by

Mr. Prathieshna Vekneswaran

Submitted in partial fulfilment of the requirements for the BSc. in Computer Science degree at the University of Westminster.

February 2024

ABSTRACT

This research project maps out the intricated field of AI for legal assistance, whereby development and implementation of the latest generation of chatbot which was designed to deal with legal inquiries within the Sri Lankan jurisprudence.

The project, built on top of pre-trained intricate language models, aims at building a domain specific dataset creation pipeline and using that dataset to train a customized chatbot which is suitable for answering legal questions largely. The implementation with advanced technologies like Llama-2-7b-chat-hf shows the power of NLP and demonstrates an effective legal knowledge retrieval method with support of technologies like PyMuPDF, Transformers, and SpaCy. This style of class compressor and LoRA adapter is a proof of eternity of efficiency and model accuracy. The chapter seal with analyses on the bot ‘s first tests, divergences found, and deficiency envisioned, finally, striving for an ideal user interface bringing about the future complex kind of hosting. As a result, capable and dedicated to solving a broad range of legal enquiries in the context of Sri Lankan Law.

Subject Descriptors:

- Computing Methodologies → Artificial Intelligence → Natural language processing→ Natural language processing
- Applied computing → Law, social and behavioral sciences → Law
- Computing Methodologies → Machine learning → Learning paradigms→ Supervised learning

Keywords: Chatbot Lawyer, Domain-specific Chatbot, Sri Lankan Legal Assistance, AI-driven Legal Research, Large Language Models, Legal Information Extraction, Dataset Creation, Transformers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Acronym	Full Form
LLM	Large Language Model
IT	Information Technology
AI	Artificial Intelligence
DL	Deep Learning

CHAPTER 01 – INTRODUCTION

1.1 Chapter Overview

The topic "LLM-Backed Chatbot Lawyer for Enhanced Legal Services in Sri Lanka" is the cornerstone of our research project. This introductory section is crucial and provides a comprehensive overview of the scope, objectives, significance, and overall structural framework of the project. Its main purpose is to give readers a clear understanding of the challenge, the project goals, the importance of our company and the sequence of the following chapters. In addition, it is tailored to explain the motivations behind this project and give readers a clear road map to navigate through the following chapters, making it easier for them to commit to the next expedition.

1.2 Problem Domain

1.2.1 Current State of Legal Services in Sri Lanka

The accessibility and affordability of legal services in Sri Lanka have been persistent concerns. According to the World Justice Project's 2020 "Rule of Law Index," Sri Lanka ranked 65th out of 128 countries in terms of accessibility and affordability of legal services. (World Justice Project, 2020) The legal system's complexity, influenced by a mix of English common law and traditional practices, presents an additional hurdle for those seeking legal assistance (Ministry of Justice and Prison Reforms, Sri Lanka, 2018). Moreover, the distribution of legal professionals is uneven, with a concentration in urban areas, leaving rural regions underserved (Sri Lanka Bar Association, 2020).

These challenges underscore the need for innovative solutions to make legal services more accessible and comprehensible to a broader range of individuals and organizations in Sri Lanka. This project aims to contribute to this effort by exploring the integration of LLM expertise into a chatbot lawyer, with the goal of addressing these pressing issues.

1.2.2 The Need for an LLM-Enhanced Chatbot Lawyer

In Lanka getting help for any legal issues or understanding the law can be a very big challenge especially for people who are in the rural areas. (Ministry of Justice and Prison Reforms, Sri Lanka, 2018) But the biggest barrier for people been the legal system been very complex to understand, due to this people get into all sort of trouble and specially people do get deceived

by others. That's where the chatbot comes into play with advanced legal knowledge the bot can offer quick affordable and reliable legal assistance specifically for Sri Lankan unique legal laws. The main goal is to make this accessible to everyone who needs it.

1.2.3 Role of Technology in Legal Services

When seen the landscape of the legal service the research has noticed that there has been transformative shift driven by technological advancement. Technology has become an indispensable tool in the world of law and order. Legal professionals use this tool to streamline their process, enhance research capabilities, and improve client interactions. The integration of AI and other automation tools has revolutionized the legal sector, has the potential to offer more efficient and accessible legal services (Smith, 2022). Legal tech innovations, including chatbots and AI-powered research tools, have shown promise in simplifying legal procedures and reducing costs while maintaining the quality of legal counsel. In the context of Sri Lanka, embracing technology in legal services has the potential to bridge the justice gap and make legal assistance more readily available to diverse segments of the population.

1.3 Problem Definition

Obtaining adequate legal aid in Sri Lanka presents considerable challenges. The country's legal system is complex and often requires basic legal knowledge to travel efficiently. This complexity and misallocation of legal professionals has resulted in expensive legal services, especially in rural areas Globally, technological advances have transformed the legal profession, creating opportunities for access to legal aid has been increased (Smith, 2022). The project aims to solve these challenges by creating an "LLM-Backed Chatbot Lawyer" equipped with advanced legal skills. The intent of this chatbot is to provide fast, cost-effective, and customized legal guidance, and ultimately differentiate the legal services available to individuals and organizations in Sri Lanka and improve the overall legal experience.

1.3.1 Problem Statement

Current lawyer chatbot are for other developed countries which has a significantly different laws compared to Sri Lanka and in order to train this chatbot using a LLM this would need an existing dataset but as per the research there no any specific dataset to train this model that's where will take all existing books, pdfs, journal, case verdicts to create domain specific chatbot

this research does not only contribute for legal advice but helps businesses and other organization to create domain specific chatbots

1.4 Aims and Objectives

1.4.1 Research Questions

RQ 1: How to collect a diverse and representative dataset of legal conversations and queries specific to Sri Lankan law?

RQ 2: How to train the chatbot to understand and generate contextually relevant responses within the specialized domain of Sri Lankan law?

RQ 3: To what extent can domain-specific language models be beneficial in enhancing the chatbot's performance within Sri Lankan law, and how should such models be trained and fine-tuned?

RQ 4: How to preprocess Law data into a dataset which the LLM model can be trained on?

1.4.2 Aims

The aim of this project is to design and develop an advanced chatbot lawyer specialized for Sri Lankan law, with a focus on addressing the existing research gap domain specific dataset creation for LLM model in providing accessible and accurate legal information to the public.

1.4.3 Research Objectives

Objective	Explanation	Learning Outcomes	Research Question
Problem Identification	**RO1 –** Aims to identify the specific challenges while developing a chatbot lawyer Sri Lankan law. **RO2 –** Recognizing areas which user often seeks legal assistance, the challenges in finding reliable information.	LO4, LO5, LO2	RQ1

Literature Review	**RO3** – Extensive examination of the existing literature, research papers, legal document and chatbot development resources to gain knowledge about chatbot technology, NLP, and legal domain. **RO4** – Systematically will review legal databases, case verdicts, journals to uncover valuable methodologies, best practices, case studies, and resources specific to Lankan law. **RO5** – Build a strong theoretical foundation towards the project allowing it to be informed about decision-making, adoption of proven strategies and avoiding pitfalls on the development	LO1	RQ1, RQ2
Data Gathering and Analysis	**RO6** – Encompassing the collection, preprocessing noisy data analyzing the data according to the LLM acceptance dataset. **RO7** – Develop technics to get the preprocess data from law books, case verdicts, legal journals. **RO8** – The objective includes creation of dataset for the legal	LO1, LO2, LO5, LO7	RQ4, RQ1

	terminology using Sri Lankan law and legal documents.		
Research Design	**RO9** – The research design cooperates for the tech stack used and chatbot architecture. **RO10** – The objective is to create a structured plan, guiding the implementation process.	LO1, LO7	RQ3, RQ4
Implementation	**RO11** – The implementation divides the work into multiple sprints dividing the development process with continuous testing. **RO12** – Deploying the API to messenger WhatsApp and other messaging applications	LO5, LO2	RQ1, RQ3, RQ4
Testing and Evaluation	**RO13** – To Evaluate the accuracy of the data which is represented by the chatbot accurate and reliable. **RO14** – Doing extrusive testing to make sure the request to the server comes accordingly and gives a response according to the request	LO4	RQ3, RQ4

Table 1 - Research Objective

1.5 Novelty of the Research

1.5.1 Problem Novelty

The Chatbot Lawyer deals with a peculiar problem in the field of legal help by analyzing the problems associated with current legal information retrieval and understanding mechanisms, particularly in the case of Sri Lanka. Traditional processes of legal research are often cumbersome and complicated for those who require specific and helpful information. Recognizing this gap in accessibility and comprehension, the project identifies a distinctive problem: the lack of customized and effective tools to help people extract meaningful legal conclusions from Sri Lankan legal documents, which limits the public's understanding and ability to navigate the legal terrain.

1.5.2 Solution Novelty

To address the arising problem, the Chatbot Lawyer proposes an innovative approach that combines modern language models, including Large Language Models (LLMs), with a customized pipeline for building a domain-specific dataset. This solution makes possible the conversion of legal documents represented in PDF documents into a structured machine-readable dataset that emphasizes the need to create a domain-specific dataset creation process. The novelty of this project is that it integrates two issues dataset creation and chatbot integration, which make the technology relevant to the peculiarities of the Sri Lankan law. Using machine learning methods, the project aims to harness an intelligent and convenient legal aid to build a user-friendly environment. However, this unique solution makes the Chatbot Lawyer stand out as an innovative approach to meeting the specific hurdles posed by the Sri Lankan legal system.

1.6 Research Gap

The research gap for this project centers on the availability and maintenance of a comprehensive and dynamically updated legal knowledge base specifically tailored to the intricacies of Sri Lankan law. (Schwarcz and Choi, 2023) This challenge encompasses several critical facets: firstly, the need for legal data to be accessible in a suitable format, spanning statutes, regulations, case law, and legal precedents; secondly, addressing the dynamic nature of the legal landscape, (Martínez, 2023) requiring mechanisms for continuous updates and

dynamic learning to ensure the chatbot's responses remain accurate and current; thirdly, achieving relevance and precision in responses by training the model using Sri Lankan legal content, preventing the chatbot from generating irrelevant information; and finally, incorporating verdicts and judgments from past Sri Lankan legal cases into the chatbot's training data, providing practical insights into the application of the law—a facet currently underrepresented in existing legal chatbot models. Addressing these gaps is crucial for the successful development of a precise, context-aware, and legally sound chatbot, ultimately enhancing access to justice and legal information in Sri Lanka.

1.7 Contribution to the Body of Knowledge

1.7.1 Contribution to the Problem Domain

- **Continuous Updates Mechanism:** To address the dynamic nature of legal content, we devise and implement mechanisms for continuous updates and dynamic learning. This ensures that the chatbot remains current with the latest legal developments, enhancing its accuracy and relevance.
- **Contextual Relevance and Precision:** The research focuses on refining the chatbot's training process to ensure that responses are contextually relevant, precise, and legally sound. This prevents the chatbot from offering extraneous or inaccurate information and enhances its utility as a legal resource.
- **Incorporation of Verdicts:** It make a pioneering contribution by incorporating verdicts and judgments from past Sri Lankan legal cases into the chatbot's training data. This enriches the chatbot's knowledge base and provides users with practical insights into the application of the law.
- **Collaborative Interdisciplinary Effort:** The project encourages collaboration between legal experts, AI researchers, and data scientists, fostering a cross-disciplinary approach to building a reliable and effective legal chatbot for Sri Lanka.
- **Improved Access to Justice:** Ultimately, our contributions aims to enhance access to justice and legal information in Sri Lanka by providing a trustworthy and user-friendly resource for individuals seeking legal guidance.

1.7.2 Contribution to Research Domain

- **Development of a Specialized Training Dataset:** The research makes a fundamental contribution by creating a specialized training dataset tailored to the unique characteristics and requirements of Specific Domain. This dataset fills a critical gap in the research domain, providing valuable resources for training chatbots to operate effectively within this specific domain. (Schwarcz and Choi, 2023)
- **Expansion of Chatbot Capabilities:** By offering a domain-specific training dataset, it can expand the capabilities of chatbots in the broader research domain. Researchers and developers can leverage this dataset to train chatbots that can comprehend, engage, and provide informed responses within Specific Domain, ultimately enhancing the utility of chatbot technology across various domains. (Shalaby et al., 2020)
- **Improved Chatbot Performance:** The availability of a domain-specific training dataset elevates chatbot performance within Specific Domain. This contribution extends to researchers and practitioners seeking to create chatbots for specific domains, providing a model for enhancing chatbot accuracy and context-awareness. (Hacker, Engel and Mauer, 2023)
- **Enhanced User Experiences:** Ultimately, the research seeks to enhance user experiences within Specific Domain by enabling the creation of chatbots that can provide more meaningful, relevant, and accurate responses. This improvement in user experiences is a valuable contribution to the broader domain of human-computer interaction. (Li, Zhang and He, no date)

1.8 Research Challenge

The research project focused on developing the Chatbot Lawyer Sri Lanka within the realm of any domain, encountered a significant research challenge. This challenge centered around the creation of a specialized training dataset that truly reflects the intricacies and nuances of the legal landscape in Sri Lanka. (Adamopoulou and Moussiades, 2020) The grappled with the task of collecting a wide array of genuine legal conversations, questions, and interactions, each of which needed to be painstakingly annotated to serve as a reliable foundation for training chatbot. Additionally, faced complex task of ensuring that the chatbot's responses not only met the stringent legal standards of Sri Lanka but were also contextually relevant and tailored to the precise needs of the users. (Baidoo-Anu and Ansah, no date) This challenge, while demanding, was vital in not only shaping the development of the Chatbot Lawyer Sri Lanka

but also in advancing the understanding of specialized language and context in the field of natural language processing, offering significant insights for the enhancement of chatbot technology in Sri Lankan law. (Firdaus, Saputra and Suprianto, 2020)

1.9 Chapter Summary

This chapter explores the capabilities of Chatbot Lawyer Sri Lanka in deepening users' comprehension of Sri Lankan law. It highlights the adaptability of the LLM to tackle domain-specific issues, moving beyond generic problem-solving. Key areas of focus include dataset creation, LM training, continuous learning, user interaction design, and ethical compliance. This chapter underscores the chatbot's potential as an educational resource, enhancing users' understanding of Sri Lankan law within specific domains.

CHAPTER 02 - SYSTEM REQUIREMENTS SPECIFICATION

2.1. Chapter Overview

This chapter is a very important guide that presents functional and non-functional requirements. By examining functional intricacies, it describes the capabilities of language comprehension, dataset usage, and answer production. Simultaneously, non-functional aspects such as performance, security and scalability are highlighted. The roadmap is going to give us an in-depth exploration of each requirement category that emphasizes its need for setting up an effective secure context-aware system.

2.2. Rich Picture Diagram

The image of affluence drawing contains the active movement within a system. It starts with data entering a pipeline where a domain specific dataset is being crafted. This data and dataset are then checked meticulously by legal experts to ensure that it is accurate and reliable. Business entities use this pipeline to make datasets that are customized according to their needs. Afterwards, that dataset goes through Falcon model smoothly resulting in a product of an intelligent chat-bot capable of understanding context. Lastly, final users interact with the chatbot; hence closing the loop in dynamic ecosystem of data generation, validation, and user interaction.

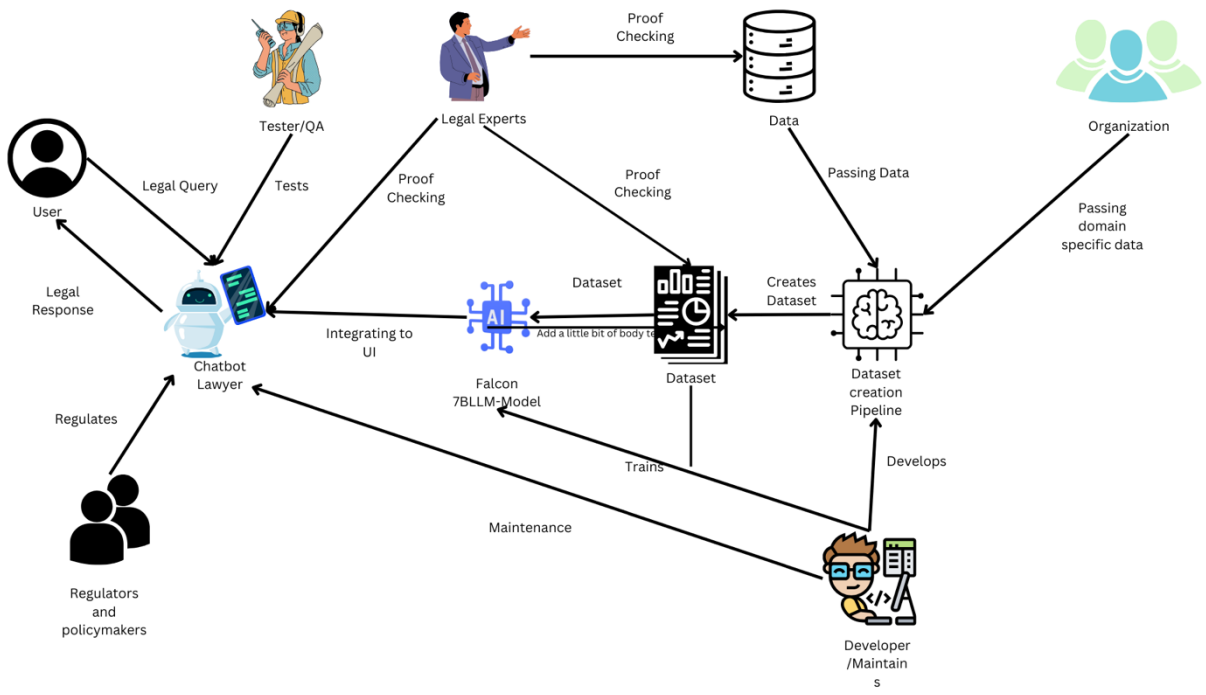


Figure 1 Rich Picture Diagram

2.3. Stakeholder Analysis

The primary stakeholders involved in the proposed prototype system, as identified in the rich-picture diagram above, are illustrated in the onion-model diagram below. Additionally, a comprehensive description of each stakeholder is provided in tabular format for improved clarity and understanding.

2.3.1. Stakeholder Onion Model

Figure 2 - Stakeholder Onion Model

2.3.2. Stakeholder Viewpoints

Stakeholder	Role	Role Description
Core System		
System Admin	System Administration	Manages and ensures the optimal performance and security of the chatbot lawyer system.
3rd Party Developers	Development	Utilizes the domain-specific dataset creation pipeline to generate customized datasets for creating their own domain-specific chatbots.
AI/ML Researchers	Research and Innovation	Conducts ongoing research to enhance the chatbot's comprehension and responsiveness in Sri Lankan legal contexts.
Containing System		
Product Owner	Product Management	Drives the development roadmap, defining features and prioritizing requirements tailored to Sri Lankan legal scenarios.
Legal Experts	Legal Compliance	Ensures the chatbot aligns with Sri Lankan legal standards, providing accurate and compliant legal information.
Public	User Interaction	Engages with the chatbot to obtain legal advice and information relevant to Sri Lanka's legal landscape.
Wider Environment		
Government Regulatory Bodies	Regulation	Oversee and enforce compliance with legal and ethical standards, ensuring the chatbot operates within regulatory boundaries.
Hacker	Security Threat	Poses a potential risk to the chatbot's security and data integrity, necessitating robust protective measures. And can get valuable data for the domain specific dataset creation pipeline of other domains
Society	Social Impact	Represents the broader societal context, reflecting cultural and legal dynamics influencing and influenced by the chatbot's usage.

Table 2 - Stakeholder Analysis

2.4. Selection of Requirements Elicitation Methodologies

Technique 1 - Literature Review
By analyzing the literature that is related, you can understand the best practices in your industry and any issues that might arise while implementing it. This approach helps to align with industry standards and importantly utilizes knowledge from previous studies in effective requirement elicitation.
Technique 2 – Structured Interviews
Structured interviews are a systematic way of engaging with stakeholders. Through well-designed queries, project team manages to access finer details regarding user needs, expectations, and preferences. This method ensures consistency during data collection as well as allows for an extensive exploration of individual views leading to a whole understanding of requirements. This method also ensures how valid the data which is been inputted and relevant to the bot or pipeline
Technique 3 - Prototyping
Prototyping creates a visual and functional model that can be evaluated by stakeholders thereby facilitating requirements elicitation. With an interactive prototype, users can get touchy-feely experience on what is being discussed. Through iterative prototyping, changing requirements may be effectively captured through this process.
Technique 4 - Use Case Analysis
In analyzing the use cases, it involves systematically critically assessing various scenarios of situations under which the machine will be used. The deer code-based approach conditions are described as being detailed user-system interactions that point at the most crucial functionality and possible challenges. The project team also gathers priceless information on practice which is necessary as a precondition for the system to operate properly, through compiling detailed use cases. In this case, there are assurances of focusing one user-system interactions while enabling the ability to capture subtle needs that arise during actual use cases.

Table 3 - Requirement Elicitation Methods

2.5. Discussion of Findings

2.5.1. Literature Review

Citation	**Findings**

(Martínez, 2023)	The study could benefit a great deal by using GPT-4's results on the Bar Test as a point of reference. The results underscore the methodological challenges that inevitably come hand in hand with the efforts aimed at determining the capabilities of the AI models and strongly advocate for the need for transparent capability evaluation. The idea found in this insight contributes to the integrity research methodology that can be employed, guaranteeing the precision of the analysis of the Chatbot Lawyer project. Since it gains an understanding of the intricacies and the tendency of the overinflation observed while inspecting GPT-4's performance, the strategic path for the research being conducted is established, relying on the creation of a legal chatbot custom-made with reference to the specifics of the Sri Lankan law. The comparative analysis with GPT-4 study's results in enriching the study in aspects of methodological approaches with practical implication and challenge behind deploying AI model in legal setting has developed.
(Bansal, 2021)	The way IBM Watson develops domain-specific chatbots is of particular interest from the perspective of our research, which deals with customer service software. By offering practical and focused insights into the useful development of domain-specific chatbots capturing the spirit of an optimal creation of a Sri Lankan. Chatbot Lawyer, IBM Watson closely adheres to the vision. Consequently, knowing IBM Watson's perspective deepens the research methodology, showing how domain-specific chatbots are built. The project learns helpful feedback concerning the improvement of large language models and their adaptation to the particularities of Sri Lankan law, which can be drawn based on the IBM Watson methodology. This comparative analysis helps in developing a contextualized and specialized legal chatbot which in turn increases the efficacy of the Sri Lankan legal chatbot in fulfilling the needs of the Sri Lankan legal system.
(Schwarcz and Choi, 2023)	This provides actionable and informed instructions to legal counselors and law students, guiding them on the benefits of properly employing

	AI large language models (LLMs) like GPT-4, Bing Chat, and Bard related to legal study and writing. The focus of this paper revolves around GPT-4, which was the most advanced LLM that was available at the time this manuscript was written, to enable lawyers to use their traditional courtroom skills at improving and validating such LLM-generated legal analyses. This paper also emphasizes the role of freely available LLMs, which can be seen as a perfect example of playing the part of highly efficient personal legal assistants to lawyers and law students.
(Kapočiūtė-Dzikienė, 2020)	As Law Fiction Theory, it is generally presumed that, once its variables apply, everyone is to be deemed aware of the law, regardless on how else one is to be considered ignorant of it laws, and therefore can be punishable by the law. This situation on the other hand has seen to the development of several cases because of a general senselessness and lack of real knowledge on the law in the society. The present paper therefore presents a plausible solution to this problem by introducing the use of chatbot platforms as possible remedies. It is therefore, meant to build chatbots that are customized to provide information to those who are looking for information regarding relevant laws. Upon asking a question concerning any aspect of the relevant documents m the administration of the United States, the bot finds searching's according to queries regarding legal documents. There is variety of command application which helps the bot in mimicking human habits thus it delivers information that the user seeks. The experimental findings indeed serve to support the chatbot's ability to identify and respond to each user correctly.

Table 4 - Elicited Requirements from Literature Review

2.5.2. Structured Interviews

Refer **APPENDIX** for interview transcripts.

Codes	**Themes**	**Analysis**
Problem Domain	Lack of efficient legal information access	The interviewee highlighted challenges in efficiently

		accessing and understanding legal information, underscoring the need for improved solutions such as the Chatbot Lawyer.
	Limited availability of domain-specific legal datasets	The interviewee pointed out the scarcity of domain-specific legal datasets, emphasizing the need for a solution like Domain-Specific Dataset Creation.
Research Gap	Absence of personalized legal assistance	The interviewee identified the absence of personalized legal assistance tools, indicating a research gap that the Chatbot Lawyer project aims to address.
	Need for tailored legal datasets	Users expressed a need for legal datasets tailored to specific domains, underscoring the importance of Domain-Specific Dataset Creation.
Methodology	User preference for AI-driven legal assistance	Users preferred AI-driven processes to enhance legal assistance, indicating a positive reception for the Chatbot Lawyer.
	Emphasis on user-friendly dataset creation	Interviewees expressed satisfaction with a user-friendly process for creating domain-specific legal datasets, enhancing the

		dataset creation methodology.
Implementation	Recognition of legal knowledge challenges	The interviewees acknowledged challenges in accessing and applying legal knowledge, emphasizing the potential impact of the Chatbot Lawyer's implementation.
User Pain Points	Limited legal knowledge access	Interviewees confirmed challenges in accessing and comprehending legal knowledge, indicating a pain point addressed by the Chatbot Lawyer.
Security Concerns	Privacy in legal queries	Users emphasized the importance of privacy in legal queries, highlighting the need for robust security measures in the Chatbot Lawyer project.
Monitoring Requirements	Feedback mechanisms for legal accuracy	Users expressed preferences for feedback mechanisms to ensure legal accuracy in the Chatbot Lawyer, emphasizing the importance of monitoring requirements.
Dataset Configurability	Ease of domain-specific dataset creation	Users prioritized a user-friendly process for creating domain-specific legal datasets, enhancing the configurability of the dataset creation methodology.

Table 5 - Elicited Requirements from Structured Interviews

2.5.3. Prototyping

Type of Prototyping	**Findings**
Throwaway/Rapid Prototyping	The rapid prototyping approach for the chatbot lawyer is mainly useful in performing initial search of design ideas and applications. However, it is effective in refining user requirements, but it is significant to address the perception and limitations associated with these disposable prototypes. First detection of potential problems improves decision-making in the latter stages of development.
Evolutionary Prototyping	The evolutionary prototyping approach makes the initial chatbot prototype evolutionary, meaning that rapid evolutionary improvements can be made to said prototype based on user insights. The one of the key features of the project is constant user participation which leads to the perfect final product that meets the customers' needs. Nevertheless, lengthier resolutions of evolutionary prototyping may possibly enlarge the duration, and sustaining continuous variations puts forth codebase amount of stability concerns.
Extreme Prototyping	The extreme supporting of Chatbot Lawyer makes notes of swift delivery of a minimum functional model with the core features. This model enables to validate the core set of functionalities within a short period of time and collects early users' feedback. But, further debugging or successive versions

	may be required with more features being added, which can affect the overall time frame allotted to the development.

Table 6 - Elicited Requirements from Prototyping

2.5.4. Use Case Analysis

The use case analysis for Chatbot Lawyer is approached by considering an elaborate interactions analysis between users and the chatbot. It pertains to different scenarios that are described using the speech act technology to explain how users will interact with the chatbot to complete certain tasks regarding legal inquiries. Analysis is used as a preliminary blueprint to understand the system function and to instantiate the chatbot response to final input. As the project begins to identify user interactions in segments and use cases, a major clarity is gained as it becomes clear that different aspects of legal information and guidance are sought by individual users in different ways. This comprehensive analysis provides a platform for the design of robust and person-friendly functionality that is incorporated in the chatbot resulting to the ability to respond to a wide scope of legal issues and situation.

2.5.5. Summary of Findings

	Finding	**Literature Review**	**Structured Interview**	**Prototyping**	**Use Case Analysis**
01	Identification of Key Features	✓	✓		✓
02	Clarification of Domain-Specific Needs	✓	✓		✓
03	User Interaction Scenarios		✓		✓
04	Validation of Domain-Specific Dataset Creation Requirements				✓
05	Early Detection of Potential Issues			✓	
06	User-Centric Functionalities		✓	✓	✓
07	Prioritization of Features	✓	✓		✓

08	Optimal Dataset Structure				✓
09	Enhanced Contextual Awareness	✓	✓		✓

Table 7 - Summary of Findings on the Elicited Requirements

2.6. Context Diagram

The context diagram below this is to reflect the followings: boundaries in Chabot lawyers and the domain specific dataset creation, and the interactions between the external entities and the flow of data among primary users.

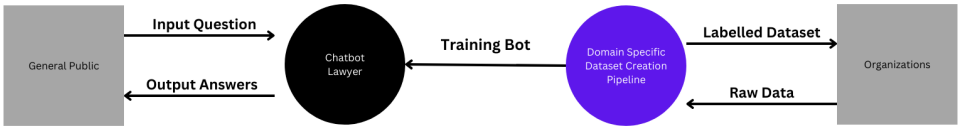


Figure 3 - Context Diagram

2.7. Use Case Diagram

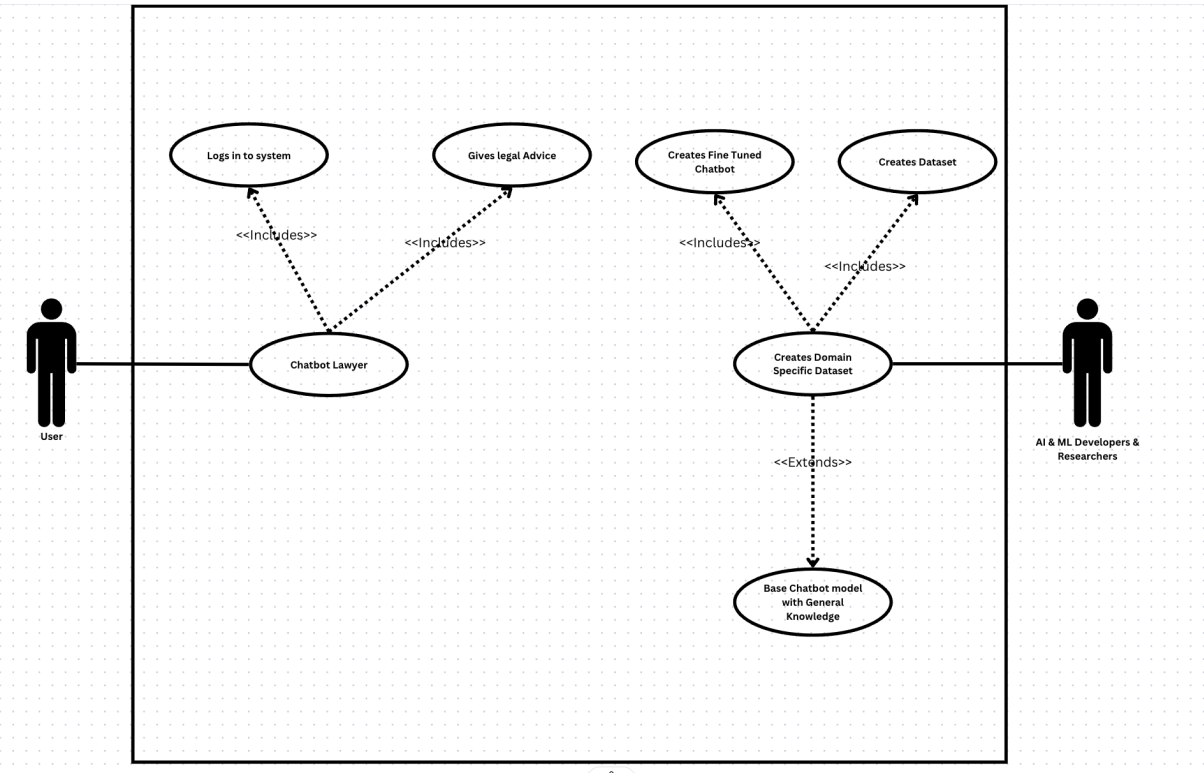


Figure 4 - Use Case Diagram

2.8. Use Case Descriptions

Use Case Name	**Chatbot Lawyer**
Description	Interacting with users, providing legal information, creating a domain-specific dataset, and assisting organizations.
Participating Actors	General Users
Stakeholders and Interests	1. Users seeking legal information. 2. Organizations utilizing the domain-specific dataset
Precondition	Chatbot Lawyer is operational, and the user initiates an interaction.
Postcondition	Legal information provided, dataset created, and organizations assisted.
Main Success Scenario	1. User interacts with the chatbot. 2. Chatbot provides legal information.

	1. Pipeline creates a domain-specific dataset. 2. Organizations utilize the dataset for chatbot creation.

Table 8 - Use Case Description for Chatbot lawyer

Use Case Name	**Create Domain Specific Dataset**
Description	Updating the Chatbot Lawyer's knowledge base, refining its responses, improving its overall performance through training, and creating a domain-specific dataset.
Participating Actors	AI & ML Developers & Researchers
Stakeholders and Interests	1. AI & ML Developers seeking model improvement. 2. Researchers ensuring the accuracy and relevance of legal knowledge. 3. Stakeholders interested in the creation of a domain-specific dataset
Precondition	The Chatbot Lawyer is in a trainable state, and the AI & ML Developer & Researcher initiates the training process.
Postcondition	Chatbot Lawyer's knowledge base is updated, a domain-specific dataset is created, and the chatbot is ready to provide more accurate and refined legal information.
Main Success Scenario	1. AI & ML Developer & Researcher initiates the training process. 2. The Chatbot Lawyer accesses relevant legal data. 3. A domain-specific dataset is created based on the accessed legal data during the training process. 4. The model is fine-tuned using the newly created domain-specific dataset. 5. The updated model and dataset are deployed for use, enhancing the chatbot's legal knowledge.

Table 9 - Use Case Description for Domain Specific Dataset Creation Model

2.9. Requirements

Priority Level	**Description**

M	Must Have	Mandatory features that were identified to be integral to the project
S	Should Have	Important features that add a significant value but aren't mandatory
C	Could Have	Supplementary features that add value and shall be implemented at the luxury of time
W	Wouldn't Have	Irrelevant or unimportant features that can be omitted altogether

Table 10 - Priority Levels of Identified Requirements

2.9.1. Functional Requirements

	Description	**Priority Level**	**Status**
FR1	Extract legal data from PDF documents	(M)	Done
FR2	Generate domain-specific questions using Machine Learning	(M)	Done
FR3	Create domain-specific dataset	(M)	Done
FR4	Get the pretrained Large Language Model up and running	(M)	Done
FR5	Fine-tune the model with Sri Lankan legal Knowledge dataset	Should Have (S)	Done
FR6	Implement chatbot interface for user queries	Should Have (S)	Pending
FR7	Provide accurate and contextually relevant legal information	Must Have (M)	Pending
FR8	Ensure security and privacy of user input	Could Have (C)	Pending
FR9	Allow integration with third-party applications	Could Have (C)	Pending

Table 11 - Functional Requirements

2.9.2. Non-Functional Requirements

	Specification	**Requirement Description**	**Status**
NFR1	Scalability	The system must be capable of handling a minimum of 500	Important

		concurrent users to accommodate potential peak usage scenarios.	
NFR2	Data Security	User inputs needs to be secure and encrypted	Important
NFR3	User Interface (UI) Design	The user interface must be designed to be intuitive and user-friendly, enhancing user interactions and overall satisfaction.	Desirable
NFR4	Compatibility	The chatbot must be compatible with major web browsers to ensure a consistent and accessible experience across different platforms.	Desirable
NFR5	Privacy Compliance	Before letting user interact with the system make sure the user accepts all terms and condition and be warned this cannot be taken as any sort of legal advice, please check on lawyer before taking any sort of action	Important
NFR6	Accessibility	The chatbot interface must be designed to be accessible to users with disabilities, following accessibility standards for inclusivity.	Desirable

Table 12 - Non-Functional Requirements

2.10. Chapter Summary

System requirements specifications thoroughly details fundamental processes that Chatbot Lawyer and, consequently, this must have to be successful. It not only describes a clear problem domain but also deals with the issue of the requirement for establishing a strong pipeline to produce a domain-specific dataset based on legal documents in the pdf format. The chapter reiterates the aspect of integration of machine learning methods notably a Large Language Model (LLM) for use in the chatbot creation. A concise list of functional and non-functional requirements sets a solid basis for further requirements development. The main objective is forming the following summary that underlines that while outlining the project's goals, its

stated that detailed and practice requirements for its dataset creation and chatbot development is essential.

CHAPTER 03 – DESIGN

3.1. Chapter Overview

This section will cover the most fascinating and tech-savvy parts of the Chatbot Lawyer and the Domain-Specific Dataset Creation. This stage is critical as the abstract design concepts find alternative and concrete systems that are based on the criteria of the project. We delve into the architectural map of the Chatbot Lawyer, giving details of the use of large language models (LLMs) as well as the bias toward certain legal domains. Furthermore, the chapter also reminds us about the systematic approach applied in developing the created domain-specific data set for the Sri Lankan legal system. Right from informational flow in the chatbot to a highly rigorous process of making data sets, the approach in this is based on coordinated and well thought out design choices. Moreover, among the critical items that we consider are user relationship, the arrangement and the nature of data and legal knowledge. The objective is to explore how both Chatbot Lawyer architecture and Domain-Specific Dataset Creation are purpose-built with the characteristics of the legal domain in mind and provide customers with a user-friendly experience as well as intelligent performance.

3.2. Design Goals

Design Goal	**Description**
User-Friendly Interface	Strive for an intuitive and user-friendly interface, ensuring that users, including legal professionals and the public, can interact effortlessly with the Chatbot Lawyer.
Legal Domain Expertise Integration	Seamlessly integrate legal domain expertise into the Chatbot, allowing for accurate and contextually relevant responses to legal queries.
Efficient Information Retrieval	Optimize information retrieval mechanisms to enable the Chatbot to access relevant legal information swiftly and accurately from the domain-specific dataset.
Context-Aware Responses	Develop the Chatbot to provide context-aware responses, considering the specifics of legal queries and nuances within the Sri Lankan legal landscape.

Robust Dataset Structuring	Implement a robust structure for the domain-specific dataset, ensuring that legal information is organized, tagged, and indexed appropriately for effective utilization.
Accessibility and Inclusivity	Prioritize accessibility, making the Chatbot and dataset accessible to users with varying levels of legal expertise, including those with limited legal knowledge.
Scalability and Future Expansion	Design the Chatbot and dataset creation system with scalability in mind, allowing for future expansion to accommodate additional legal domains or jurisdictions.

Table 13 - Design Goals

3.3. High Level Design / System Architecture Design

The high-level design, or system architecture design, for the Chatbot Lawyer and domain-specific dataset creation project is intricately crafted to provide an overarching structure that seamlessly integrates legal assistance and dataset creation functionalities. This design prioritizes modularity, scalability, and user-friendliness, ensuring an intuitive interface for both legal professionals and individuals seeking legal information. Emphasizing efficient information retrieval and context-aware responses, the architecture enables the Chatbot to navigate the nuanced landscape of Sri Lankan legal queries. Robust dataset structuring enhances accessibility, making legal information available to users with varying levels of expertise. Additionally, the design is forward-looking, poised for scalability and future expansion to accommodate evolving legal domains or jurisdictions. The high-level design serves as a guiding framework for the system, promising adaptability, and sophistication in delivering legal assistance and dataset creation services.

3.3.1. Architecture Diagram

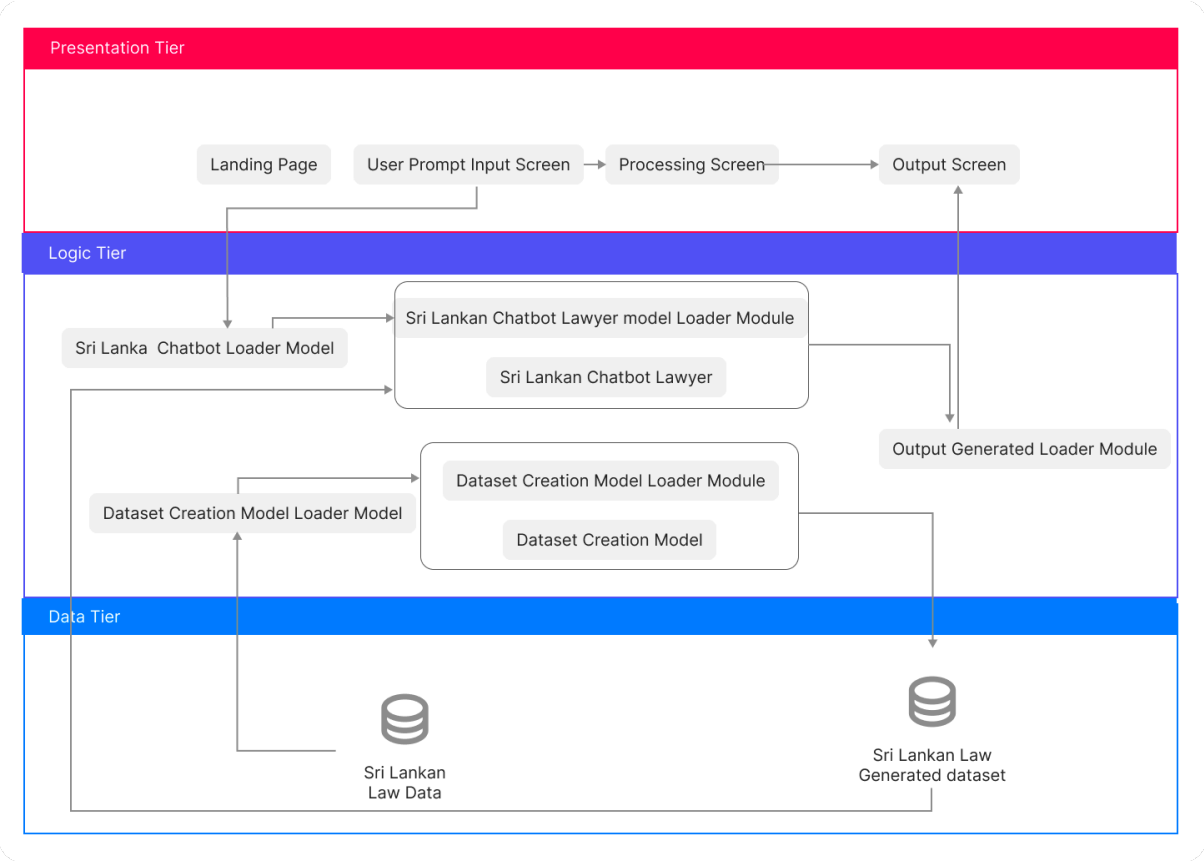


Figure 5 - High Level Architecture Diagram

3.3.2. Discussion of Tiers

Data Tier:

Sri Lankan Law Data: The data tier is the formatting of the comprehensive dataset consisting of Sri Lankan Legal data. These databases hold the law journey- a goal that only such databases can help to attain. This database provides the fundamental basis of the legal knowledge for the entire system to function, to a high level of accuracy.

Sri Lankan Law Generated Dataset: After the law data is passed into the domain specific dataset creation model a labelled dataset is created which can be used to train LLM model

Logic Tier:

Sri Lankan Chatbot Loader Model: These scripts are responsible for loading the chatbot and performing its initialization. It is amongst the principal component in charge of elements putting in place together, assembling the logic, and preparing the chatbot for user interactions.

Sri Lankan Chatbot Lawyer Model Loader Module: In such a demonstration, the defined function pertains to the Chatbot Lawyer, which is responsible for model loading and initialization. Not only is this process performed to make sure that the chatbot has all the specified legal knowledge to provide precise answers but also the language is created in a way that provides an automatic authorization for chatbot to provide legal support.

Sri Lankan Chatbot Lawyer: The module has the main brain which contains all the needed intelligence. It answers the user queries, uses legal databases, and produces information fit for the task.

Dataset Creation Model Loader Module: First, we will discuss the loading of the Model for Dataset Creation which keeps the key mechanism of the model in mind. Such system acts as a sentinel or quarantine for the world's states; thereby being able to design and structure context sensitive legal datasets. (Goutham, 2020)

Dataset Creation Model: Module: customizing datasets by way of using legal knowledge and environment. It generates datasets while portraying Sri Lankan legal situations. The system which adapts and builds a better one with changing legal norms gets promoted by it.

Output Generated Loader Module: Such a module handles the important performance of the Sri Lankan Chatbot Lawyer and the Dataset Creation Models. It guarantees smooth merging of response and data flow into the front end. This is made possible through diverse kinds of media.

Presentation Tier:

Landing Page: The startup screen that invites the user to navigate the software. It helps to get started by offering you an opportunity to get the first-hand overview and consultancy on the matter or by supplying you with the proper dataset**.**

User Prompt Input Screen: This is basically chatting style screen where the prompt and output will be shown in a chat format.

Processing Screen: This is the same chat screen with the loading icon showing that the request has been sent.

Output Screen: Again the output screen is also the same chat screen with the response from the server

3.4. Low Level Design / System Design

3.4.1. Design Paradigm

The Structured Systems Analysis and Design Method (SSADM) is the paradigm chosen for the project in the making up of the Chatbot Lawyer and the domain-specific dataset creation project. This choice essentially is because the project requires a structured and phased approach whereby all issues related to building of a chatbot**,** and creation of domain specific dataset solution will be addressed beginning with the simplest. SSADM was most brilliant in the use of structural model to analyze components of computer system, as well as in reporting comprehensive documentation and keeping a logical sequence during the feature development. This technique completely reflects the communications systems and the goals of collaborative work, creating prerequisites to stipulate SSADM rather than Object-Oriented Analysis and Design (OOAD) for their suitability in systematic and efficient development3.5. Detailed Design Diagrams

3.5.1. Component Diagrams

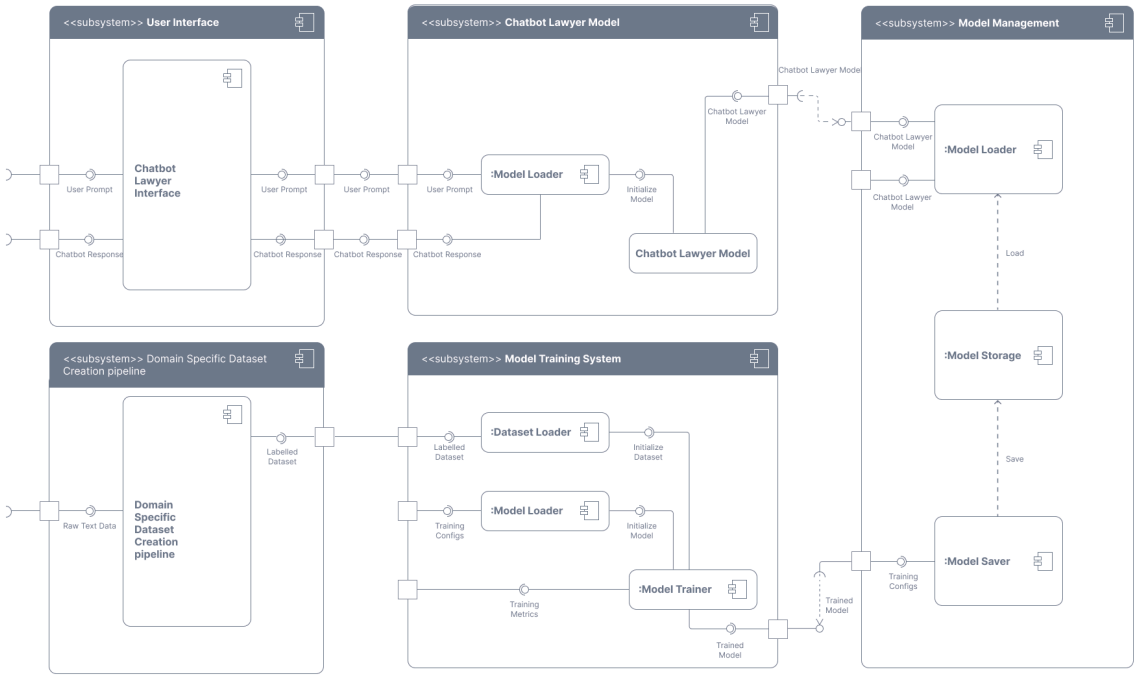


Figure 6 - Component Diagram

3.5.2. System Process Flow Chart

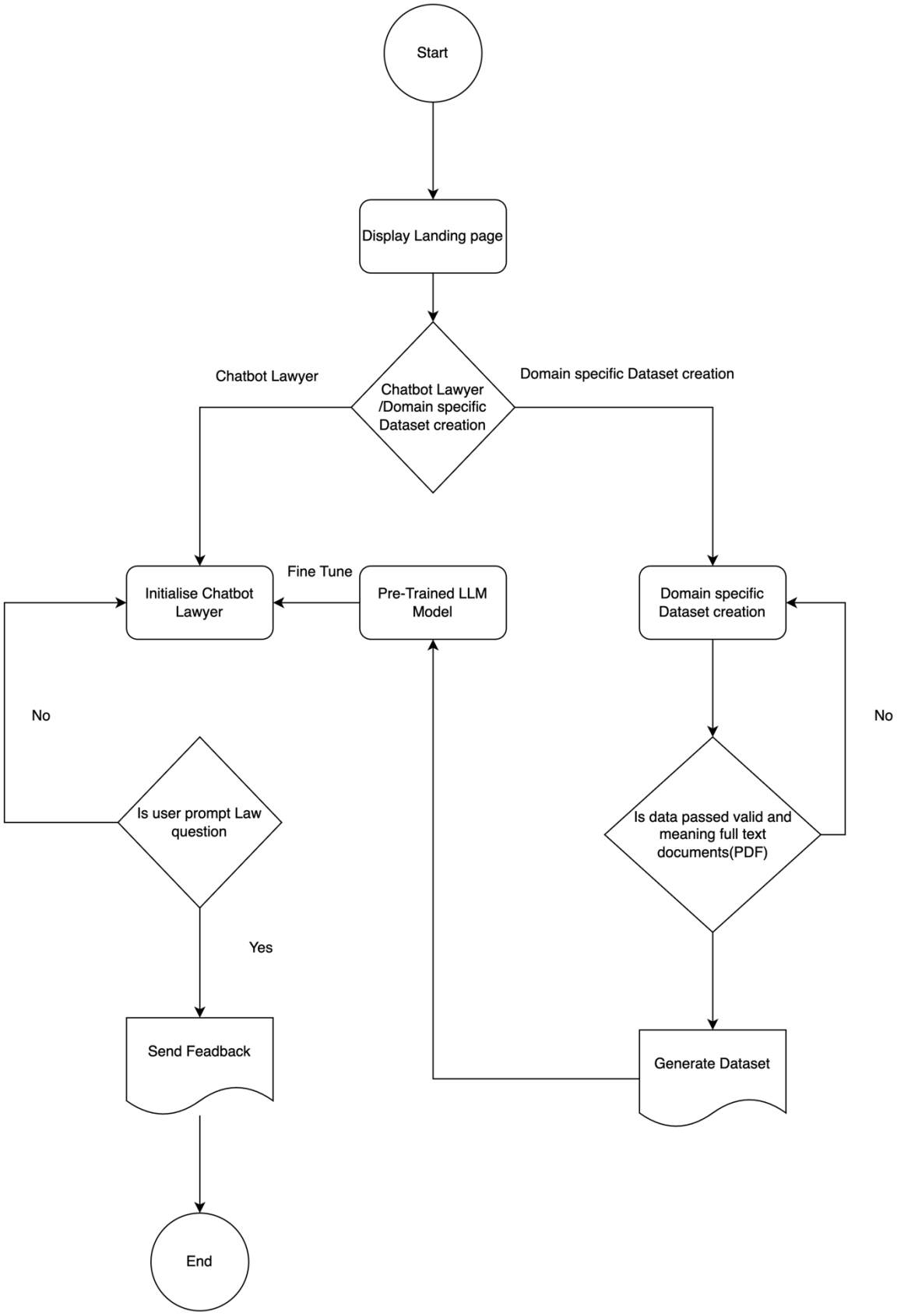


Figure 7 - Flow Chart

3.5.3. User Interface Design (Low Fidelity Wireframe)

Figure 8 - Chatbot Wireframe

3.6. Chapter Summary

AI Chatbot lawyer, as well as domain-specific dataset creation, faced the challenges of AI in legal assistance with the help of finding problem domain and conducting a systematic literature review. Through the solution of the difficulty of laborious screenshot handling and by underscoring the fact that the case law system does not possess any specialized tools, the project aims to create a gap that was left in the legal reality. The arising of a Sri Lankan legal dataset and a chatbot lawyer made in Sri Lanka were evident when we headed for a technical review. Its unique elements, which comprise the country specific approach and the server-side app, differentiate it from the other AI solutions for the law, with their prospect to develop for the Sri Lankan legislation such assistance that can be adapted to the Sri Lankan legal intricacies using the Structured Systems Analysis and Design Method (SSADM).

CHAPTER 04 – INITIAL IMPLEMENTATION

4.1. Chapter Overview

In this chapter, the focus shifts from conceptualization to practical application as the initial implementation of the Chatbot Lawyer and Domain-Specific Dataset Creation project unfolds. This stage marks the translation of design and planning into tangible outcomes. The chapter outlines the key steps taken, challenges encountered, and the iterative process of bringing the project from the drawing board to its early operational phase.

4.2. Technology Selection

4.2.1. Technology Stack

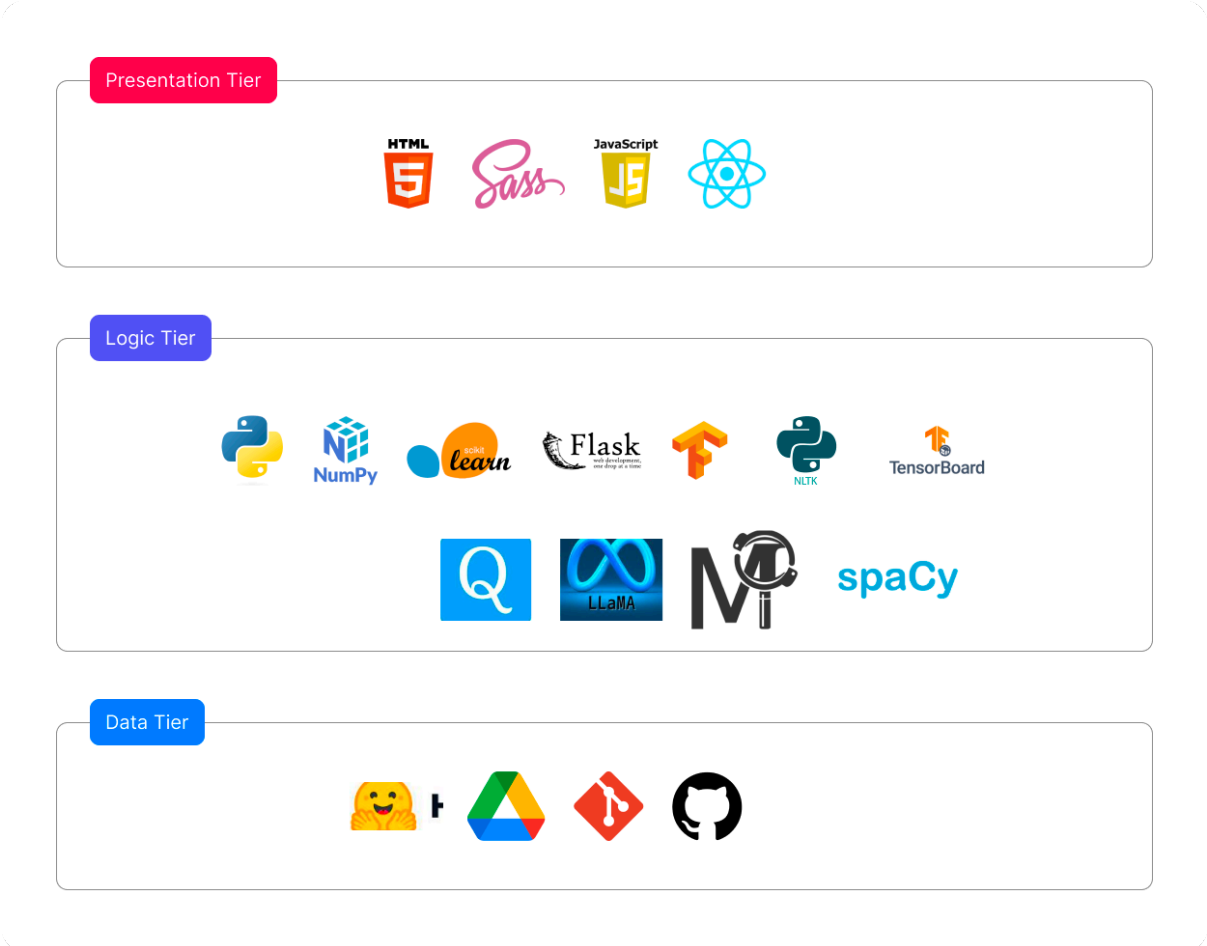


Figure 9 - Technology Stack

4.2.2. Data Set Selection

The dataset for the Chatbot Lawyer and Domain-Specific Dataset Creation project was meticulously crafted by extracting raw law data from the official website of the Parliament of Sri Lanka (parliament.lk). The process involved comprehensive scraping and collection methods to ensure a diverse and representative set of legal documents. The raw law data obtained serves as the foundation for creating a robust and contextually relevant dataset, aligning with the specific legal nuances of Sri Lanka. This strategic approach to data selection aims to enhance the effectiveness and accuracy of the chatbot in providing tailored legal information to users. (Goutham, 2020)

4.2.3. Development Frameworks

Framework	Rationale
Flask	Flask was chosen as the primary web framework for its simplicity, flexibility, and efficient development cycle.
Questgen	Questgen is a crucial framework for question generation, contributing to the dataset enrichment process.
Hugging Face Transformers	Leveraging the Transformers library from Hugging Face provides access to pre-trained language models for legal tasks.

Table 14 - Selection of Development Framework

4.2.4. Programming Languages

Programming Language	Reasoning
Python	Python was chosen as the primary programming language due to its extensive libraries, readability, and widespread use in the fields of AI, machine learning, and web development.

Table 15 - Selection of Programming Languages

4.2.5. Libraries

Library/Toolkit	Rationale
transformers	Transformers provides state-of-the-art natural language processing (NLP) models, including GPT (Generative Pre-trained Transformer), essential for language-related tasks in the project.

datasets	The datasets library facilitates efficient management and access to datasets, ensuring seamless integration with the project's requirements.
accelerate	Accelerate optimizes the performance of PyTorch computations, enhancing the efficiency of deep learning tasks and contributing to faster model training.
peft	Peft is a Persian text extraction tool that aids in processing Persian language data, supporting multilingual capabilities in the project.
trl	The TRL library is employed for Text Representation Learning, enhancing the representation, and understanding of textual data within the project.
bitsandbytes	Bitsandbytes contributes to the processing of binary data, a crucial aspect when handling various data formats in the context of the project.
fitz	Fitz, or PyMuPDF, supports PDF document manipulation, allowing for efficient extraction and processing of text from legal documents in the project.
numpy	Numpy, a fundamental library for numerical operations in Python, is upgraded to ensure compatibility and leverage the latest enhancements for efficient computation.
spaCy	SpaCy, with version 2.3.3, serves as a reliable natural language processing (NLP) library, contributing to the linguistic analysis aspects of the project.
pke	Pke, a Python Keyphrase Extraction library, is quietly installed to support the extraction of key legal phrases and terms, enhancing the project's information retrieval.
nltk	The NLTK (Natural Language Toolkit) universal tagset is downloaded to facilitate universal part-of-speech tagging, contributing to linguistic analysis in the project.

Table 16 - Selection of Libraries

4.2.6. IDE

IDE / Code Editor	Justification
Google Colab	Google Colab was chosen for its accessibility, collaborative features, and pre-installed libraries essential for the project.

	The cloud-based nature allows easy sharing and collaboration on the development of the chatbot lawyer and domain-specific dataset creation.
VS-Code	This is to create a frontend using react

Table 17 - Selection of IDEs

4.2.7. Summary of Technology Selection

Component	**Tools & Technologies**
Programming Language	Python
Development Framework	Flask
Question Genration Model	Questgen
Deep Learning Library	Transformers
Other Libraries	datasets, accelerate, peft, trl, bitsandbytes, fitz, PyMuPDF, Numpy, spaCy, pke, nltk
IDEs	Google Colab
Version Control	Git, GitHub, Hugging Face

Table 18 - Summary of Selected Technologies

4.3. Implementation of the Core Functionality

Domain Specific Dataset Creation

The domain-specific pipeline for the chatbot lawyer project is implemented with a focus on extracting relevant information from legal documents. Initially, the text is extracted from a PDF file using PyMuPDF, followed by a preprocessing step to clean the extracted text. To facilitate efficient processing, the text is chunked into manageable segments. A Question Generation (QG) model generates context-aware questions for each text chunk. These questions are then fed into a pre-trained BERT-based Question Answering (QA) model to produce answers within the provided context. The results, including questions, context, and answers, are formatted, and organized in a DataFrame, which is subsequently saved as a CSV file ('output_dataset.csv'). This pipeline lays the groundwork for incorporating advanced functionalities into the chatbot lawyer system, enhancing its capabilities in understanding, and responding to legal queries. (Goutham, 2020)

```python
import fitz  # PyMuPDF
import transformers
from transformers import AutoTokenizer,BertTokenizerFast, BertForQuestionAnswering
import torch
import pandas as pd
# Define the bert tokenizer
tokenizer = AutoTokenizer.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')

# Load the fine-tuned modeol
model = BertForQuestionAnswering.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')

def generate_answer(question, context):

  inputs = tokenizer.encode_plus(question, context, return_tensors='pt')

  outputs = model(**inputs)
  answer_start = torch.argmax(outputs[0])  # get the most likely beginning of answer with the argmax of the score
  answer_end = torch.argmax(outputs[1]) + 1

  answer = tokenizer.convert_tokens_to_string(tokenizer.convert_ids_to_tokens(inputs['input_ids'][0][answer_start:answer_end]))

  return answer

  # Encode the input text and question, and get the scores for each word in the text


# Find the words in the text that corresponds to the highest start and end scores
# with a torch.no_grad():
#     outputs = model(**inputs)


# Function to extract text from a PDF file
def extract_text_from_pdf(pdf_file):
    doc = fitz.open(pdf_file)
    text = ""
    for page in doc:
        text += page.get_text()
    return text

# Function to chunk text into pieces of a specified size
def chunk_text(text, chunk_size=4000):
    chunks = [text[i:i + chunk_size] for i in range(0, len(text), chunk_size)]
    return chunks

# Function to preprocess text (you can customize this)
def preprocess_text(text):
    # Remove extra whitespace and special characters
    text = ' '.join(text.split())
    return text

# Load your PDF file
pdf_file_path = '/content/constitution1.pdf.zip'

# Extract text from the PDF
pdf_text = extract_text_from_pdf(pdf_file_path)

# Preprocess the text
cleaned_text = preprocess_text(pdf_text)

# Chunk the text into 2000-token pieces
text_chunks = chunk_text(cleaned_text, chunk_size=512)

payload = {"input_text": ""}
question_context_pairs = []  # Store question-context pairs
i = 1
# Initialize a list to store the generated answers with context
answers_with_context = []
```

```python
# Loop through chunks and make predictions
for chunk in text_chunks:
    payload["input_text"] = chunk
    outputs = qg.predict_shortq(payload)

    # Check the structure of the outputs dictionary
    if 'questions' in outputs:
        question_list = outputs['questions']
    elif 'your_custom_key' in outputs:
        question_list = outputs['your_custom_key']
    else:
        # Handle the case when the structure is different
        print("Unexpected structure in the 'outputs' dictionary. Check the structure and update the code.")
        continue

    # Iterate through the extracted questions and contexts
    for item in question_list:
        question = item.get('Question', '')  # Use get to avoid KeyError
        context = item.get('context', '')  # Use get to avoid KeyError

        # Check if question and context are non-empty before processing
        if question and context:
            question_context_pairs.append((question, context))
            answer = generate_answer(question, context)

            # Include context in the instruction field
            instruction_text = f"You are a problem-solving assistant. Before answering, explain your reasoning step-by-step. Context: {context}\nFinal answer: {answer}"

            answers_with_context.append({
                "input": f"Question: {question}\nContext: {context}",
                "output": answer,
                "instruction": instruction_text
            })

# Convert the list of answers with context to a DataFrame
answers_df = pd.DataFrame(answers_with_context)

# Save the DataFrame to your desired format (e.g., CSV)
answers_df.to_csv('output_dataset.csv', index=False)

# Print the generated answers with context
print(answers_df)
```

Chatbot Lawyer

The provided code represents the initial implementation of the chatbot lawyer, focusing on training and evaluation. The model is based on the Llama-2 architecture, fine-tuned on a dataset containing Sri Lankan legal information obtained from 'zoom12/SriLankaLaw.' This is the dataset which has been created from the domain specific dataset creation pipeline .The pipeline involves loading the base model, configuring quantization settings, and implementing Lora adapters for training efficiency. The model is trained using a customized TrainingArguments setup, leveraging the BitsAndBytesConfig for quantization. The training process incorporates paged AdamW optimization and linear learning rate scheduling. Supervised fine-tuning is facilitated by the SFTTrainer, considering a maximum sequence length of 512 tokens. The resulting trained model is saved as 'llama-2-7b-Sri-Lankan-Law,' and a text generation pipeline is applied to respond to a given legal query. (Touvron et al., 2023)

```
[ ]  #Model
    base_model = "NousResearch/Llama-2-7b-chat-hf"
    #Fine-tune model name
    new_model = "llama-2-7b-Sri-Lankan-Law"
    #Load the Dataset from hugging face
    dataset = load_dataset("zoom12/SriLankaLaw", split="train")
    #Tokenizer
    #Load the tokenizer from Llama 2
    tokenizer = AutoTokenizer.from_pretrained(base_model, use_fast=True)
    #In Llama2 we dont have the padding token which is a very big problem, because we have a dataset with different number of tokens in each row.
    #So, we need to pad it so they all have the same length and here i am using end of sentence token and this will have an impact on the generation of our model
    #I am using End of Sentence token for fine-tuning
    tokenizer.pad_token=tokenizer.eos_token
    tokenizer.padding_side="right"
```

```
[ ]  #Configration of QLoRA
    #Quantization Configuration
    #To reduce the VRAM usage we will load the model in 4 bit precision and we will do quantization
    bnb_config = BitsAndBytesConfig(
        load_in_4bit=True,
        #Quant type
        #We will use the "nf4" format this was introduced in the QLoRA paper
        bnb_4bit_quant_type="nf4",
        #As the model weights are stored using 4 bits and when we want to compute its only going to use 16 bits so we have more accuracy
        bnb_4bit_compute_dtype=torch.float16,
        #Quantization parameters are quantized
        bnb_4bit_use_double_quant=True,
    )


    # LoRA configuration
    peft_config = LoraConfig(
        #Alpha is the strength of the adapters. In LoRA, instead of training all the weights, we will add some adapters in some layers and we will only
        #train the added weights
        #We can merge these adapters in some layers in a very weak way using very low value of alpha (using very little weight) or using a high value of alpha
        #(using a big weight)
        #15 is very big weight, usually 32 is considered as the standard value for this parameter
        lora_alpha=15,
        #10% dropout
        lora_dropout=0.1,
        bias="none",
        task_type="CAUSAL_LM",
    )

    # Load base moodel
    model = AutoModelForCausalLM.from_pretrained(
        base_model,
        quantization_config=bnb_config,
        device_map={"": 0}
    )

    model.config.use_cache = False
    model.config.pretraining_tp = 1

    # Cast the layernorm in fp32, make output embedding layer require grads, add the upcasting of the lmhead to fp32
    #prepare_model_for_kbit_training----> This function basically helps to built the best model possible
    model = prepare_model_for_kbit_training(model)
```

```
[ ] # Set training arguments
    training_arguments = TrainingArguments(
            output_dir="./results",
            num_train_epochs=1,#3.5 good for the Llama 2 Model
            per_device_train_batch_size=4,# Number of batches that we are going to take for every step
            gradient_accumulation_steps=1,
            evaluation_strategy="steps",#Not helpful because we donot want to evaluate the model we just want to train it
            eval_steps=1000,
            logging_steps=25,
            optim="paged_adamw_8bit",#Adam Optimizer we will be using but a version that is paged and in 8 bits, so it will lose less memory
            learning_rate=2e-4,
            lr_scheduler_type="linear",
            warmup_steps=10,
            report_to="tensorboard",
            max_steps=-1, # if maximum steps=2, it will stop after two steps
    )

    # Set supervised fine-tuning parameters
    trainer = SFTTrainer(
        model=model,
        train_dataset=dataset,
        eval_dataset=dataset,#No separate evaluation dataset, i am using the same dataset
        peft_config=peft_config,
        dataset_text_field="instruction",
        max_seq_length=512,# In dataset creation we put a threshold 2k for context length (input token limit) but we dont have enough VRAM unfortunately it will take a lot of VRAM to put everything into memory so we are just gonna stop at 512
        tokenizer=tokenizer,
        args=training_arguments,
    )

    # Train model
    trainer.train()

    # Save trained model
    trainer.model.save_pretrained(new_model)
```

```
# Run text generation pipeline with our model
#Input Prompt
prompt = "Is weed legal in sri lanka"
#Wrap the prompt using the right chat template
instruction = f"### Instruction:\n{prompt}\n\n### Response:\n"
#Using Pipeline from the hugging face
pipe = pipeline(task="text-generation", model=model, tokenizer=tokenizer, max_length=128)
result = pipe(instruction)
#Trim the response, remove instruction manually
print(result[0]['generated_text'][len(instruction):])

Weed is illegal under the Misuse of Drugs Ordinance No.12 of 1981 in Sri Lanka, and the country has a strict legal framework regarding drug use and possession. For the cultivation, production, manufacture, distribution, sale or advertising of any narcotic drug, the person will be li
```

4.4. Chapter Summary

"Initial implementation" chapter comprises the core of the chatbot lawyer concept development. These parts address the factors that determine the choice of the dataset, frameworks that help in creating datasets, the programming languages and libraries or toolsets that are fundamental for the dataset creation and chatbot training. The concepts of pivotal functionalities are elaborated consciously through lines of code which shows how the language models are crafted finest, the model quantization settings in the configuration of the architecture, and how the Lora architecture is used. Furthermore, the model training covered the hyperparameter selection, the choice of optimizers, and the supervised fine-tuning processes are also described. The chapter is concluded by presenting the trained model as a practical application through a text generation pipeline - chatbot serves making the discussion more precise regarding the specific legal inquiry.

CHAPTER 05 – CONCLUSION

5.1. Chapter Overview

In this section we will be concluding on what this chatbot lawyer and Domain Specific Dataset creation model will be used and how it can be improved in the future. Will be discussing the deviations from scope related deviations schedule related deviation, how was the initial test result with proper screenshots will be discussed below. Furthermore, a video demonstration will be added as a link to see how the initial implementation of this project looks.

5.2. Deviations

5.2.1. Schedule Related Deviations

Item	Project Roadmap	Actual Status
Initial Project Research Proposal	September 2023	Completed
Project Initiation Document (PID)	October 2023	Completed
Literature Review (LR)	November 2023	Completed
Software Requirements Specification (SRS)	November 2023	Completed
Proof of Concept (PoC)	December 2023	Completed
Project Specifications & Design Prototype (PSDP)	February 2024	Completed
Minimum Viable Product (MVP)	March 2024	In-Progress
Testing & Evaluation	March 2024	In-Progress
Final Thesis	April 2024	In-Progress

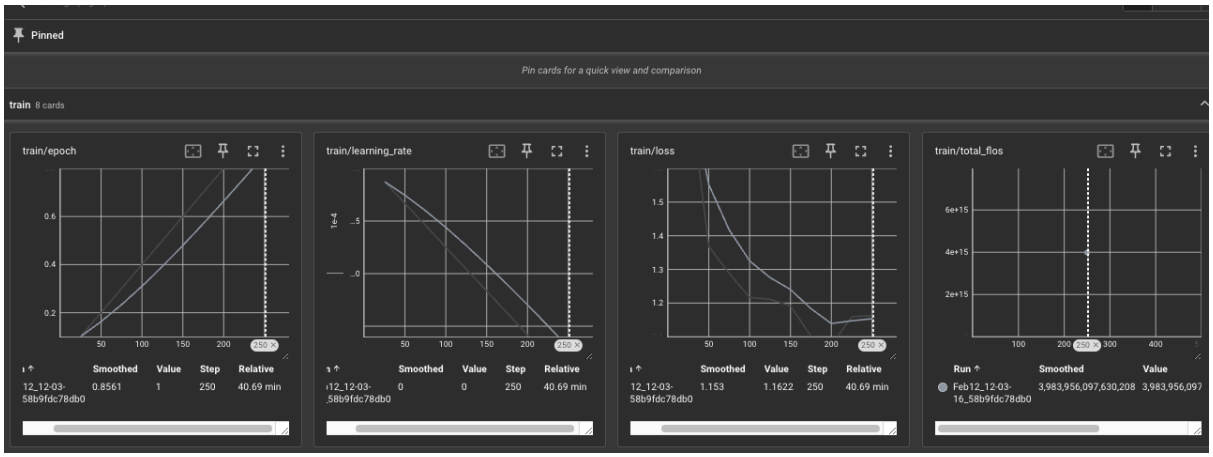
Table 20 - Schedule Related Deviations

5.3. Initial Test Results

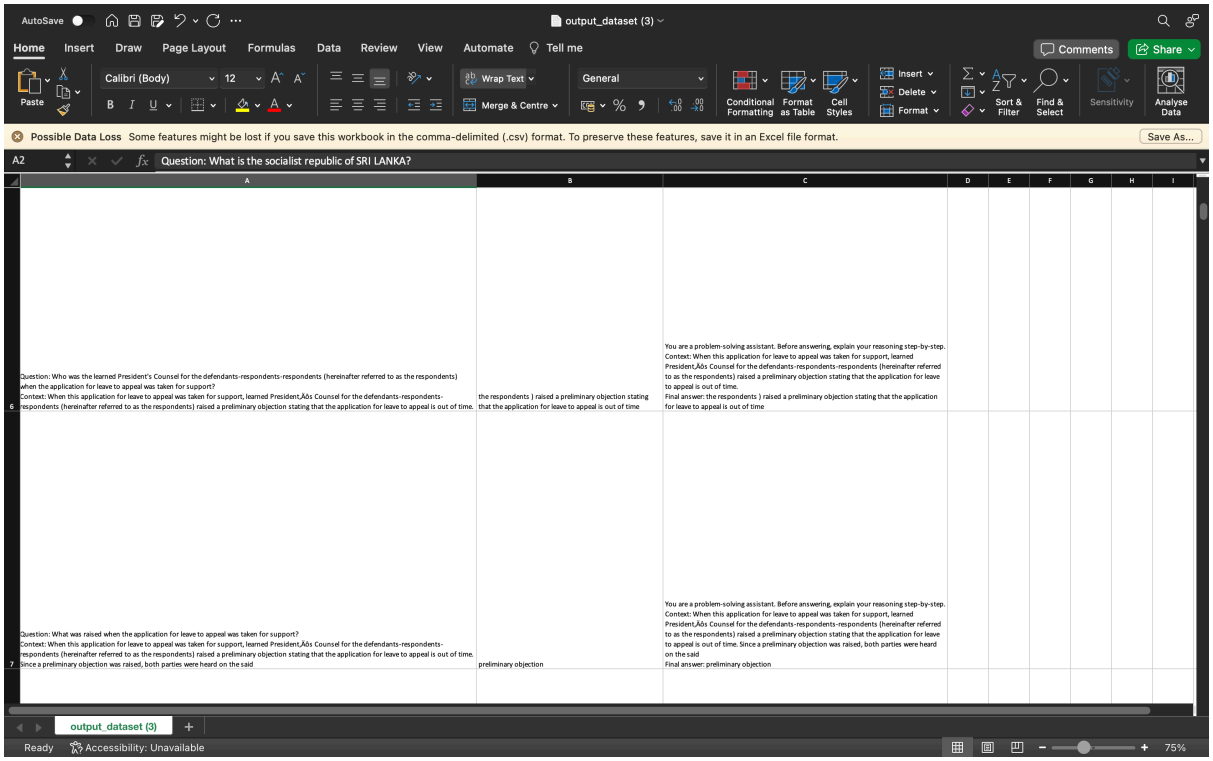
In the section on initial test results, the research project delves into the outcomes of preliminary testing, presenting findings that will be substantiated with accompanying screenshots. This segment offers a snapshot of the project's performance at an early stage, providing a glimpse

into the functionality and effectiveness of the implemented systems. By incorporating visual evidence through screenshots, the discussion aims to bolster the credibility of the presented results and offer readers a tangible glimpse into the project's initial testing phase.

In the below screenshot can see how the fine-tuning training performance was using TensorBoard



Generated Dataset output CSV file



Chatbot lawyer Output from Fine Tune model

5.4. Required Improvements

the project identifies areas for enhancement, emphasizing two key aspects: GUI (graphical user interface) improvements and chatbot implementation of the project. This conversation will be about the requirement of a user-friendly GUI which makes it easier for users to use this system and make all their activities streamlined. Besides that, technical components of bringing the chatbot into public view are also looked at as I am yet to make clear that it's public. Through considering these facets, project leads the way for future development, and guarantees the wider usability of the product due to its ideal state.

5.5. Demo of Prototype

A Demonstration can be found on this link for the project.

https://youtu.be/dp6EboVsjQc

The code base can be accessed in this repository.

https://github.com/omarShiraz/chatbotLawyer

5.6. Chapter Summary

Within this chapter, the project has been systematically advanced to the initial phase of Implantation where work has centered on the production of the chatbot lawyer and the appointment of a domain-specific dataset. Applying modern models, like NousResearch/Llama-2-7b-chat-hf and using frameworks for example, PyMuPDF, Transformers, and SpaCy, could clearly demonstrate that an already trained BERT model is able to generate right Law answers. The application of methods, e.g. quantization and the addition of adapters (LoRA) indicate that there is a chance for model changes through which the developers will seek to ensure high performance and efficiency. The chapter ends with a brief analysis of future implications where there is scope for improvement of the GUI and the need for the chatbot to be live for public and all-inclusive access.

REFERENCES

Anwar, S. et al. (2022). Image Colorization: A Survey and Dataset. Available from http://arxiv.org/abs/2008.10774 [Accessed 1 September 2023].

Cohn, M. and Martin, R.C. (2006). *Agile estimating and planning*. Upper Saddle River, NJ: Prentice Hall Professional Technical Reference.

Huang, S. et al. (2022). Deep learning for image colorization: Current and future prospects. *Engineering Applications of Artificial Intelligence*, 114, 105006. Available from https://doi.org/10.1016/j.engappai.2022.105006.

Kim, H., Kim, Jonghyun and Kim, Joongkyu. (2022). Image-to-Image Translation for Near-Infrared Image Colorization. *2022 International Conference on Electronics, Information, and Communication (ICEIC)*. 6 February 2022. Jeju, Korea, Republic of: IEEE, 1–4. Available from https://doi.org/10.1109/ICEIC54506.2022.9748773 [Accessed 31 August 2023].

Le-Tien, T. et al. (2021). GAN-based Thermal Infrared Image Colorization for Enhancing Object Identification. *2021 International Symposium on Electrical and Electronics Engineering (ISEE)*. 15 April 2021. Ho Chi Minh, Vietnam: IEEE, 90–94. Available from https://doi.org/10.1109/ISEE51682.2021.9418801 [Accessed 1 September 2023].

Liang, W., Ding, D. and Wei, G. (2021). An improved DualGAN for near-infrared image colorization. *Infrared Physics & Technology*, 116, 103764. Available from https://doi.org/10.1016/j.infrared.2021.103764.

Liu, L. et al. (2022). Optimal LED Spectral Multiplexing for NIR2RGB Translation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022. New Orleans, LA, USA: IEEE, 12642–12650. Available from https://doi.org/10.1109/CVPR52688.2022.01232 [Accessed 12 September 2023].

Liu, Y. et al. (2023). Learning to colorize near-infrared images with limited data. *Neural Computing and Applications*, 35 (27), 19865–19884. Available from https://doi.org/10.1007/s00521-023-08768-7.

Ma, X. et al. (2022). Near-Infrared Image Colorization Using Asymmetric Codec and Pixel-Level Fusion. *Applied Sciences*, 12 (19), 10087. Available from https://doi.org/10.3390/app121910087.

Wang, F., Liu, L. and Jung, C. (2020). Deep Near Infrared Colorization with Semantic Segmentation and Transfer Learning. *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. 1 December 2020. Macau, China: IEEE, 455–458. Available from https://doi.org/10.1109/VCIP49819.2020.9301788.

Yang, Z. and Chen, Z. (2020). Learning From Paired and Unpaired Data: Alternately Trained CycleGAN for Near Infrared Image Colorization. *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. 1 December 2020. Macau, China: IEEE, 467–470. Available from https://doi.org/10.1109/VCIP49819.2020.9301791

Zeger, I. et al. (2021). Grayscale Image Colorization Methods: Overview and Evaluation. *IEEE Access*, 9, 113326–113346. Available from https://doi.org/10.1109/ACCESS.2021.3104515.

Zhang, R., Isola, P. and Efros, A.A. (2016). Colorful Image Colorization. Available from http://arxiv.org/abs/1603.08511.

Zhao, M. et al. (2022). CSTGAN: Cycle Swin Transformer GAN for Unpaired Infrared Image Colorization. *2022 3rd International Conference on Control, Robotics and Intelligent System*. 26 August 2022. Virtual Event China: ACM, 241–247. Available from https://doi.org/10.1145/3562007.3562053 .

Zhou, S. and Kamata, S.-I. (2022). Near-Infrared Image Colorization with Weighted UNet++ and Auxiliary Color Enhancement GAN. *2022 7th International Conference on Image, Vision and Computing (ICIVC)*. 26 July 2022. Xi'an, China: IEEE, 507–512. Available from https://doi.org/10.1109/ICIVC55077.2022.9887040.

Goutham, R. (2020) Questgen - An open source NLP library for Question generation algorithms, Towards Data Science. Available at: https://towardsdatascience.com/questgen-an-open-source-nlp-library-for-question-generation-algorithms-1e18067fcdc6?gi=f5de4e0a0fca

Touvron, H. et al. (2023) "Llama 2: Open foundation and fine-tuned chat models," ArXiv, abs/2307.09288. Available at: http://arxiv.org/abs/2307.09288

APPENDIX

Interview Transcripts:

Note these interviews was done physically and written after the interview has occurred.

Interview 1: Law Student Dino Arulanantham

Interviewer: Thank you for joining us today. Can you share your experiences in dealing with legal information access challenges?

Interviewee: Absolutely. In my practice, accessing and understanding legal information efficiently has been a constant challenge. There's a need for tools that can provide personalized legal assistance.

Interviewer: What are your thoughts on the availability of domain-specific legal datasets?

Interviewee: Well, the truth is, we often find ourselves lacking access to datasets tailored to specific legal domains. Having such datasets would greatly enhance our work.

Interviewer: How do you feel about the concept of an AI-driven Chatbot that provides legal assistance?

Interviewee: I think it's a game-changer. Having an AI-driven Chatbot that understands legal nuances and can assist in real-time would significantly improve our efficiency.

Interview 2: Damitha Wimalasooriya Software Engineering Manager for Research and Development at Acumatica

Interviewer: Thank you for your time. Can you share your preferences regarding AI-driven processes in legal assistance?

Interviewee: Certainly. AI-driven processes are crucial in enhancing legal assistance. They bring efficiency and accuracy to complex legal scenarios.

Interviewer: How important is user-friendliness in creating domain-specific legal datasets?

Interviewee: User-friendliness is key. Researchers and developers need a straightforward process to create datasets tailored to specific legal domains. It ensures broader usability.

Interviewer: What's your take on the implementation of the Chatbot Lawyer project?

Interviewee: Recognizing challenges in accessing legal knowledge and addressing them through the Chatbot Lawyer is a commendable step. It could revolutionize how legal information is accessed.

Interview 3: Sri Lankan Lawyer Anonymous

Interviewer: As a legal professional, your insights are crucial. What challenges do you face in accessing and researching legal information?

Interviewee: Navigating extensive legal databases can be time-consuming. The sheer volume of information makes it challenging to quickly find relevant legal precedents or updates.

Interviewer: How do you envision a Chatbot aiding in legal research?

Interviewee: A Chatbot capable of processing complex legal queries and providing concise, accurate information would be a game-changer. It could significantly expedite legal research tasks.

Interviewer: Regarding privacy and confidentiality, what concerns would you have when interacting with a Chatbot for legal queries?

Interviewee: Privacy is paramount in legal matters. Knowing that the Chatbot adheres to strict confidentiality measures and doesn't compromise sensitive information is essential for its credibility.

Interviewer: How important is user feedback to ensure the Chatbot's legal accuracy?

Interviewee: User feedback is invaluable. It serves as a constant quality check, helping refine the Chatbot's responses and ensuring it stays updated with the latest legal developments.