ARBA MINCH UNIVERSITY

ARBA MINCH UNIVERSITY INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING AND SOFTWARE

INTERNSHIP REPORT

Project Title: Amharic Legal Chat Bot

Hosting Company: Information Network Security Administration (INSA)

Name: PETER MESSAY **ID. NSR/1244/14**

Organization Mentor: **Mr. Terefe Feyisa** (Data Scientist, INSA)

Academic Supervisor: **Mrs. Sindu T**. (Lecturer, AMU)

Internship Duration: February 8, 2025 – March 15, 2025

Arba Minch, Ethiopia

Sep 25,2025

ORGANIZATION

Name: Information Network Security Administration (INSA)

Region: Addis Ababa

District: HQ Addis Ababa

City: Addis Ababa

Telephone: +251-113-71-71-14

MY MENTOR IN THE ORGANIZATION

Name: **Mr. Terefe Feyisa**

Profession: Data Scientist

Position: Data Analytics Researcher at INSA, Addis Ababa

Phone: +251-913-23-90-36

Email: terefefeyisa@gmail.com

EXAMINERS APPROVAL

We, the undersigned examiners of this internship report have evaluated and approved the Corrected version of that final report as per the guideline of Arba Minch Institute of Technology.

1) Examiner Name _____ Sign _____ Date_____

2) Examiner Name _____ Sign _____ Date_____

3) Faculty IE Faciliatory_____ Sign_____ Date_____

DECLARATION

I, Peter Messay, a fifth-year Software Engineering student at Arba Minch University (AMU), hereby affirm that the internship project titled "Amharic Legal Chatbot" is my original work, conducted at the Information Network Security Administration (INSA) from February 8, 2025, to June 15, 2025.

This report has not been submitted to any other university or institution for the purpose of obtaining a degree, diploma, or certificate. All sources of information and materials used from other references have been properly acknowledged.

Peter Messay
Date: _____

Approval by Academic Supervisor

This is to certify that the internship report titled "Amharic Legal Chatbot", submitted by Peter Messay, a fifth-year Software Engineering student at Arba Minch University, was carried out under my supervision.

I hereby recommend and approve this report for submission as partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering.

Name: _____

Academic Supervisor
Signature: _____
Date: _____

This is to certify that the internship project entitled "Amharic Legal Chatbot" was undertaken by Peter Messay at the Information Network Security Administration (INSA) under my guidance.

I confirm that the work documented in this report is the result of the intern's effort and commitment during the internship period from February 8, 2025, to June 15, 2025.

Name: _____
Organization Mentor
Signature: _____
Date: _____

ACKNOWLEDGMENTS

First and foremost, I am deeply grateful to Almighty God for blessing me with health, wisdom, and strength throughout the course of my internship.

My sincere appreciation goes to the Information Network Security Administration (INSA) for granting me the opportunity to complete my internship in such a professional and innovative environment. I would like to extend my heartfelt thanks to my mentor, Mr. Terefe Feyisa, whose continuous guidance, encouragement, and insightful feedback were invaluable in enriching my learning experience and ensuring the successful completion of my project.

I am also profoundly thankful to my academic supervisor, Mrs. Sindu T. of Arba Minch University, for her consistent support, constructive advice, and close supervision during the preparation of this report.

A special note of gratitude goes to my beloved family—my mother, father, and sister—for their unconditional love, financial support, and unwavering encouragement. Their sacrifices and motivation have been the foundation of my academic journey.

I am equally grateful to my friends and dormmates in Addis Ababa, who supported me like family during my stay. Despite the challenges of high living costs, they generously shared what they had, cared for me wholeheartedly, and stood by me in countless ways. Their companionship made this journey more meaningful and fulfilling.

Lastly, I wish to acknowledge all individuals who, in one way or another, contributed to the success of my internship and the preparation of this report. To all of you, I extend my heartfelt gratitude.

ACRONYMS

Acronym	Full Form
AI	Artificial Intelligence
INSA	Information Network Security Administration
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
LLM	Large Language Model
API	Application Programming Interface
UI/UX	User Interface / User Experience
CSV	Comma-Separated Values
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
EDA	Exploratory Data Analysis
ML	Machine Learning

EXECUTIVE SUMMARY

This report presents an overview of my internship at the Information Network Security Administration (INSA), conducted from February 8, 2025, to June 15, 2025. The internship provided practical exposure to Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP), with a special emphasis on Amharic language applications.

The primary focus of my internship was the development of an Amharic Legal Chatbot, designed to deliver accessible and reliable legal information in Amharic. The system was implemented using a Retrieval-Augmented Generation (RAG) framework, where legal documents were embedded and stored in a vector database. This architecture enabled efficient semantic search and context-aware responses, ensuring accurate and user-friendly interactions.

Alongside the chatbot, I also worked on a Sentiment Analysis sub-project involving text classification. This effort enhanced my understanding of core NLP techniques, including tokenization, feature extraction, and supervised learning approaches.

To support these projects, I participated in structured training programs that strengthened my skills in Python programming, data preprocessing, visualization, algorithm optimization, and model evaluation. I also gained hands-on experience in building end-to-end ML pipelines, deploying models in real-world contexts, and collaborating effectively within a professional team setting.

In addition to technical expertise, the internship improved my soft skills in communication, teamwork, time management, and problem-solving. Most importantly, it highlighted the transformative potential of AI in addressing challenges faced by low-resource languages such as Amharic.

Overall, this internship not only provided a strong foundation for my future career in software engineering but also reinforced my passion for leveraging AI to solve real-world problems.

Table of Contents

CHAPTER ONE

1. ORGANIZATION PROFILE OF INFORMATION NETWORK SECURITY ADMINISTRATION (INSA)

1.1 BRIEF HISTORY

The Information Network Security Administration (INSA) is Ethiopia's national body responsible for safeguarding cybersecurity and protecting critical information infrastructure. It was first established in 1999 under Council of Ministers Regulation No. 130/1999, with the mandate of securing the country's information and communication networks against emerging digital threats. Although Ethiopia's adoption of digital technologies was still in its early stages, the government recognized the potential risks posed by cyber threats to national security, economic growth, and governance.

As technology advanced and cybercrime grew more complex, INSA underwent a series of legal and organizational reforms. In 2003, it was restructured under Regulation No. 250/2003, which broadened its responsibilities in information protection and cyber defense. Later, Proclamation No. 808/2006 further strengthened its authority to address the rising challenges of international cybercrime, digital espionage, and threats targeting Ethiopia's critical infrastructures.

In 2021, through the Definition of Powers and Duties of the Executive Organs of the Federal Democratic Republic of Ethiopia (Proclamation No. 1263/2014), the institution was formally renamed the Information Network Security Administration. This restructuring reflected Ethiopia's commitment to adapting its cybersecurity strategy to contemporary technological and geopolitical realities.

Today, INSA plays a pivotal role in Ethiopia's digital transformation, focusing on building national cybersecurity capacity, offering advisory services to government agencies, securing digital platforms, and defending the country against cyberattacks. Its vision is "to realize national capability that ensures information superiority," while its mission is "to protect the national interest through building a capacity that enables safeguarding the country's infrastructures." INSA also upholds core values such as resilience, integrity, respect for people and the law, and a commitment to making a difference.

1.2 MAIN PRODUCTS AND SERVICES

Since its establishment, INSA has evolved into a multi-faceted cybersecurity institution providing services and solutions that protect, defend, and advance Ethiopia's digital security. Its products and services can be broadly categorized into the following:

1.2.1 CYBERSECURITY AND NETWORK PROTECTION

- INSA's core service is the protection of Ethiopia's information and communication infrastructure.
- It provides **cyber defense mechanisms** against hacking, malware, phishing, and other cyberattacks.
- It also ensures the **resilience of national data centers, communication networks, and government IT systems**.

1.2.2 CYBER THREAT INTELLIGENCE AND INCIDENT RESPONSE

- INSA collects and analyzes cyber threat intelligence to detect vulnerabilities and forecast possible attacks.
- It operates **Computer Emergency Response Teams (CERTs)** that provide incident detection, analysis, mitigation, and recovery services to both government and public institutions.

1.2.3 DIGITAL FORENSICS AND CYBERCRIME INVESTIGATION

- INSA supports law enforcement and judiciary bodies through **digital forensic investigations**, recovering and analyzing electronic evidence from cybercrime cases.
- It provides expert services in cyber law enforcement and contributes to the development of national policies on cybercrime prevention.

1.2.4 CRYPTOGRAPHY AND SECURE COMMUNICATION SYSTEMS

- To ensure **confidentiality, integrity, and availability** of sensitive information, INSA develops and deploys **encryption technologies** and secure communication platforms.
- These systems are widely used by government institutions for secure data exchange and classified communications.

1.2.5 CAPACITY BUILDING AND TRAINING

- INSA conducts professional **training programs in cybersecurity, digital forensics, cryptography, and network security**.
- It also partners with universities and research institutions to build local capacity in advanced technologies such as **Artificial Intelligence (AI), Machine Learning, and Big Data analytics**.

1.2.6 POLICY DEVELOPMENT AND CYBERSECURITY REGULATIONS

- INSA plays a crucial role in drafting **national cybersecurity policies, standards, and regulations**.
- It ensures that government institutions and private sectors comply with cybersecurity frameworks that safeguard Ethiopia's digital assets.

1.2.7 RESEARCH AND DEVELOPMENT (R&D)

- INSA invests in research on emerging technologies such as **AI, cloud computing, blockchain, and natural language processing (NLP)**.
- Through R&D, it develops innovative solutions tailored to Ethiopia's needs, including **local-language AI applications**, like the Amharic Legal Chatbot project in which I participated.

1.2.8 PUBLIC AWARENESS AND DIGITAL SAFETY CAMPAIGNS

- INSA also provides **awareness programs** to educate the general public and organizations about safe digital practices.
- It regularly launches campaigns to protect citizens from online fraud, misinformation, and privacy violations.

1.3 MAIN CUSTOMERS AND END USERS

The **Information Network Security Administration (INSA)** serves a wide range of customers and stakeholders, both directly and indirectly. Its core mission of safeguarding Ethiopia's information infrastructures means that its services are vital to government bodies, public institutions, private organizations, and even individual citizens. The following are the major customers and end users of INSA's products and services:

1.3.1 GOVERNMENT INSTITUTIONS

Government ministries, agencies, and regional state offices are INSA's primary customers. These institutions rely on INSA for:

- **Cybersecurity protection** of their networks, websites, and data centers.
- **Secure communication platforms** to safeguard sensitive state information.
- **Digital forensics services** in case of data breaches or cybercrime investigations.
- **Policy guidance** to ensure compliance with national cybersecurity regulations.

Examples include the **Ministry of Defense, Ministry of Foreign Affairs, Ministry of Finance, and law enforcement agencies**, all of which handle critical and sensitive data that require high levels of security.

1.3.2 CRITICAL INFRASTRUCTURE PROVIDERS

Organizations that manage Ethiopia's **critical infrastructures** such as telecommunications, energy, banking, aviation, and transport are also key customers of INSA.

- Telecommunications companies rely on INSA to mitigate cyber threats targeting networks and data transmission.
- Financial institutions use its advisory and monitoring services to protect against **cyber fraud, digital theft, and ransomware attacks**.
- Power and energy companies benefit from INSA's resilience-building services that secure industrial control systems against cyber sabotage.

1.3.3 JUDICIARY AND LAW ENFORCEMENT AGENCIES

The judiciary and police departments use INSA's **digital forensics** and **cybercrime investigation services** to gather electronic evidence in criminal cases. INSA acts as a technical partner in prosecuting cyber-related crimes, ensuring that evidence collected is both reliable and admissible in court.

1.3.4 PRIVATE SECTOR ORGANIZATIONS

Private companies, particularly those in the **banking, insurance, ICT, and e-commerce sectors**, benefit from INSA's services in the following ways:

- **Consultancy and training** on cybersecurity best practices.
- **Incident response and recovery support** in case of attacks.

- **Access to secure communication technologies** for protecting customer data and business secrets.

This relationship ensures that the private sector is resilient against the growing risks of cyberattacks while fostering trust among customers and stakeholders.

1.3.5 ACADEMIC AND RESEARCH INSTITUTIONS

Universities and research centers are important customers of INSA's **capacity-building and training programs**. INSA collaborates with higher education institutions to:

- Train students and researchers in **cybersecurity, AI, and advanced computing techniques**.
- Provide **internship and research opportunities** for university students (such as my internship experience).
- Jointly develop **innovative solutions** in areas like natural language processing, cryptography, and cyber policy.

1.3.6 GENERAL PUBLIC AND CITIZENS

While INSA primarily serves organizations, its services indirectly reach Ethiopia's general population. Ordinary citizens are beneficiaries of INSA's efforts through:

- **Public awareness campaigns** on safe internet use, digital privacy, and online fraud prevention.
- **Cybersecurity guidelines** that help protect personal devices and data.
- **Secure national digital platforms** that enhance public trust in e-government services.

For instance, when citizens use online government portals, mobile banking, or digital communication platforms, the safety of their information is indirectly ensured by INSA's systems.

1.3.7 INTERNATIONAL PARTNERS

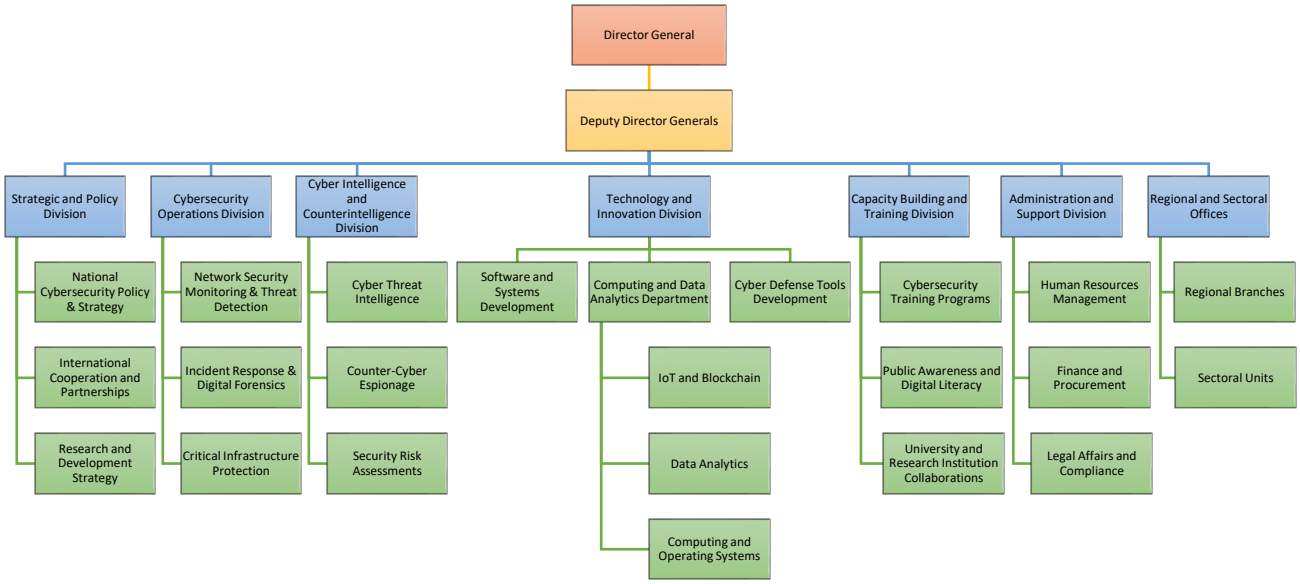
In an interconnected world, cybersecurity is not confined to national borders. INSA collaborates with **regional and international organizations** in areas such as:

- Cyber threat intelligence sharing.
- Joint training programs on advanced cyber defense.

- Research collaborations in AI, big data, and digital security.

Through these partnerships, INSA also positions Ethiopia as a responsible actor in the global cybersecurity arena.

1.4 ORGANIZATIONAL STRUCTURE OF INSA



The **Information Network Security Administration (INSA)** is Ethiopia's national cybersecurity agency mandated to secure the country's digital infrastructure, safeguard critical information systems, and promote technological innovation. Its organizational structure integrates **top leadership**, **technical divisions**, and **administrative support units** to carry out its strategic objectives.

Top Leadership

- **Director General** – Provides overall leadership, strategy, and policy direction.
- **Deputy Director Generals** – Support the Director General by managing specific operational areas and divisions.

Strategic and Policy Division

- National Cybersecurity Policy & Strategy
- International Cooperation and Partnerships
- Research and Development Strategy

Cybersecurity Operations Division

- Network Security Monitoring & Threat Detection
- Incident Response & Digital Forensics
- Critical Infrastructure Protection

Cyber Intelligence and Counterintelligence Division

- Cyber Threat Intelligence
- Counter-Cyber Espionage
- Security Risk Assessments

Technology and Innovation Division

- **Software and Systems Development** – Secure platforms and enterprise systems.
- **Cyber Defense Tools Development** – Indigenous cybersecurity solutions.
- **Computing and Data Analytics Department**
    - IoT and Blockchain – Research and implementation.
    - Data Analytics (AI & Machine Learning) – Advanced analytics and AI research.
    - Computing and Operating Systems – Secure computing architectures.

Capacity Building and Training Division

- Cybersecurity Training Programs
- Public Awareness and Digital Literacy
- University and Research Institution Collaborations

Administration and Support Division

- Human Resources Management
- Finance and Procurement
- Legal Affairs and Compliance

Regional and Sectoral Offices

- Regional Branches – Localized operations and outreach.
- Sectoral Units – Support for defense, finance, telecommunications, and other critical industries.

1.5 WORK FLOW OF INSA

Workflow of INSA

The workflow of the Information Network Security Administration (INSA) reflects a **hierarchical but collaborative system**, where strategic leadership directs policies, technical divisions implement operations, and supporting units ensure smooth execution. Below is a structured explanation of how the workflow proceeds within INSA:

1.5.1 STRATEGIC LEADERSHIP AND POLICY DIRECTION

- The **Director General** and **Deputy Directors General** initiate the workflow by defining the **national cybersecurity priorities**, policy frameworks, and long-term strategies.
- Strategic goals are developed in alignment with Ethiopia's national security, digital transformation, and technology advancement objectives.
- These directions are communicated to all divisions for implementation.

1.5.2 STRATEGIC AND POLICY DIVISION – PLANNING AND COORDINATION

- This division **translates leadership priorities into actionable policies and programs**.
- It develops **cybersecurity strategies**, coordinates with **international partners**, and establishes **research agendas**.
- The division also ensures that all other technical divisions operate under a unified policy framework.
- Output from this stage flows to both **operations divisions** and **innovation units**.

1.5.3 CYBERSECURITY OPERATIONS DIVISION – REAL-TIME SECURITY WORKFLOWS

- Receives directives and frameworks from the **Strategic and Policy Division**.
- Executes **continuous monitoring of networks**, **threat detection**, and **incident response**.
- Coordinates with the **Cyber Intelligence Division** for threat intelligence.
- When threats are detected, the division escalates cases to **Incident Response & Forensics Teams**, who investigate and mitigate risks.
- Results and reports are fed back to leadership for decision-making and long-term adjustments.

1.5.4 CYBER INTELLIGENCE AND COUNTERINTELLIGENCE DIVISION – INTELLIGENCE WORKFLOW

- Gathers **cyber threat intelligence** from both internal monitoring and external sources.
- Performs **risk assessments** to identify vulnerabilities in critical infrastructures.
- Works with **operations teams** to provide intelligence support during live incidents.
- Shares analysis with **policy and strategy teams** to adapt future plans.
- Counter-espionage teams monitor for infiltration, insider threats, and foreign cyber interference.

1.5.5 TECHNOLOGY AND INNOVATION DIVISION – RESEARCH AND DEVELOPMENT WORKFLOW

- Receives **strategic priorities** from leadership and **technical requirements** from operations teams.
- Develops **indigenous tools**, including secure software, cyber defense platforms, and AI/ML systems.
- The **Data Analytics Department** builds advanced **AI solutions**, performs **data-driven research**, and enhances **semantic search systems**.
- Innovations are tested, validated, and transferred to **operations teams** for real-world deployment.

- Feedback loop ensures that R&D responds directly to current national cyber needs.

1.5.6 CAPACITY BUILDING AND TRAINING DIVISION – KNOWLEDGE AND SKILLS WORKFLOW

- Based on gaps identified by **operations** and **intelligence units**, this division designs **training programs**.
- Conducts **public awareness campaigns** to improve digital literacy among citizens.
- Collaborates with **universities and research institutions** for skill development.
- Ensures continuous **capacity building** to sustain Ethiopia's cyber defense ecosystem.

1.5.7 ADMINISTRATION AND SUPPORT DIVISION – ORGANIZATIONAL SUPPORT WORKFLOW

- Handles **HR recruitment**, **finance**, and **legal compliance**.
- Provides the **logistical and administrative backbone** for all technical divisions.
- Ensures that technical operations, innovation, and intelligence units have the **resources and legal clearances** needed for effective functioning.

1.5.8 REGIONAL AND SECTORAL OFFICES – DECENTRALIZED WORKFLOW

- Act as **local extensions** of INSA in regional states and key economic sectors.
- Implement national cybersecurity strategies at **regional levels**.
- Provide **sector-specific support** for industries like defense, finance, and telecommunications.
- Report back to headquarters on local challenges, creating a feedback loop for national strategy.

CHAPTER TWO

2. ABOUT PROJECT

Title: Amharic Legal Chatbot

Amharic Legal Chatbot: An AI-Powered Platform for Accessible Legal Information

2.1 PROJECT SUMMARY

During my internship, I worked on an innovative project titled **Amharic Legal Chatbot**, designed to provide Ethiopian citizens with reliable and easily accessible legal information in their native language. The main objective was to build a platform where individuals could conveniently obtain answers to their legal inquiries from home, thereby enhancing legal literacy and supporting the modernization of information access in Ethiopia.

The project's core emphasis was on creating a **user-friendly online platform** that enables citizens to access accurate legal information quickly and efficiently. By digitizing legal knowledge, the chatbot helps minimize the need for in-person visits to legal offices, reducing both time and financial costs. This approach is especially beneficial for people living in rural and underserved areas, where access to legal services is often limited.

Moreover, the initiative aligned with the broader mission of the Information Network Security Administration (INSA) to promote Amharic language technologies and foster secure, innovative digital solutions for Ethiopia. Through INSA's support and collaboration, I was able to strengthen my technical skills while contributing to a project with tangible societal benefits.

2.1.1 EXECUTIVE SUMMARY

The technical backbone of the Amharic Legal Chatbot was a **Retrieval-Augmented Generation (RAG) framework**, which combines semantic search with generative AI techniques to deliver factual, context-aware, and safe responses. A structured workflow was followed to implement this system:

1. **Data Collection** – Amharic legal documents were gathered from official and authentic sources.
2. **Data Preprocessing** – he collected materials were cleaned, normalized, and segmented for consistency..

3. **Chunking and Embedding** – Texts were divided into smaller units and converted into vector embeddings using transformer-based models optimized for the Amharic language.
4. **Vector Database Storage** – These embeddings were stored in a vector database, enabling fast and precise semantic retrieval.
5. **System Integration –** A Flask-based web interface was created to provide users with a simple and visually appealing platform for interacting with the chatbot.

This end-to-end workflow, carried out in **Google Colab** and integrated with modern NLP tools, ensured the chatbot's ability to understand Amharic queries and return accurate legal information.

2.1.2 EXPECTED OUTPUT

The anticipated outcome of the project was a **fully functional Amharic Legal Chatbot** capable of both retrieving and generating accurate Amharic responses to legal queries. By allowing citizens to communicate in their native language, the chatbot guarantees **inclusivity and accessibility**.

The tool was designed for **multiple user groups**—not only the general public but also law students, researchers, and legal professionals—who could leverage it to quickly access statutes, references, and case-related materials, thereby improving the **efficiency of legal research and decision-making.**

Furthermore, by minimizing dependence on physical consultations and enabling **on-demand access to verified legal resources**, the chatbot supports the goals of **Digital Ethiopia 2025** and contributes to advancing AI-driven solutions for low-resource languages. Ultimately, the Amharic Legal Chatbot bridges the gap between legal knowledge and public access, empowering citizens while reinforcing Ethiopia's **digital sovereignty**.

2.2 PROBLEM STATEMENT AND JUSTIFICATION

2.2.1 PROBLEM STATEMENT

Accessing **Ethiopian legal information** continues to be a major challenge due to both **linguistic barriers** and **technical limitations**. While most legal documents—including **proclamations, regulations, and directives**—are written in Amharic, they are often published in **lengthy, unstructured formats** that are difficult for citizens to interpret. This lack of

organization forces **citizens, law students, and legal professionals** to spend considerable time manually searching through dense texts to identify relevant legal provisions.

Existing **global AI tools** such as **ChatGPT** or **Google Gemini** are not well-suited for Amharic and lack proper coverage of **Ethiopian-specific legal frameworks**. As a result, they cannot provide **accurate or dependable assistance** to Amharic-speaking users.

Traditional methods of obtaining legal advice make the situation worse. Citizens frequently need to **visit legal offices in person** or **consult experts**, a process that is both **time-consuming and expensive**. For individuals in **remote or rural regions**, access to such resources is even more limited, leading to **systemic inequality in legal awareness and access to justice**.

Consequently, there is a **critical knowledge gap** between the availability of legal documents and their **actual accessibility** for the public. Without innovative interventions, many Ethiopian citizens will remain **uninformed about their rights and responsibilities**, limiting their ability to make **well-informed decisions**.

2.2.2 JUSTIFICATION

The development of the **Amharic Legal Chatbot** is both timely and necessary to address the challenges outlined above. The justification for this project can be summarized as follows:

- **Bridging the Language Gap**: Most AI-based legal retrieval tools are designed for English and other dominant global languages, leaving Amharic-speaking users excluded. A chatbot that fully supports Amharic ensures **linguistic inclusivity** and **cultural relevance**.
- **Enhancing Legal Awareness**: Legal texts are often lengthy, technical, and difficult to interpret. By **simplifying and structuring information**, the chatbot enables citizens to better understand their **rights and obligations**, thereby improving **legal literacy**.
- **Supporting Professionals and Students**: **Law students, researchers, and practitioners** need quick and reliable access to legal resources. The chatbot streamlines the search process, reducing the time spent on manual lookups and enhancing **academic and professional productivity**.
- **Advancing Digital Transformation**: The project is aligned with **Digital Ethiopia 2025**, a national strategy emphasizing digitization and equitable access to services. By making legal information available online, the chatbot contributes directly to Ethiopia's **digital transformation agenda**.
- **Driving Innovation for Low-Resource Languages**: Amharic is classified as a **low-resource language** in NLP research. Developing this chatbot provides valuable insights

and tools that can also be applied to other **under-resourced languages**, expanding the global AI knowledge base.

- **Reducing Physical Barriers**: With online access to legal information, citizens no longer need to make frequent visits to legal institutions. This reduces **financial costs and time burdens**, especially for individuals in **rural or underserved communities**.
- **Empowering Society**: Ultimately, the chatbot fosters **legal empowerment**, enabling citizens to make **informed decisions**, building **trust in digital systems**, and supporting Ethiopia's vision of becoming a **knowledge-driven society**.

2.3 OBJECTIVES OF THE PROJECT

2.3.1 GENERAL OBJECTIVE

The overall objective of this project was to design and implement an AI-powered Amharic Legal Chatbot that delivers accessible, reliable, and user-friendly legal information to citizens, students, and professionals in their native language.

2.3.2 SPECIFIC OBJECTIVES

To accomplish the above goal, the project was guided by the following specific objectives:

- **To collect and preprocess Amharic legal documents** (proclamations, regulations, and directives) from authentic and official sources.
- **To design a knowledge representation system** by converting legal texts into embeddings and storing them in a vector database for efficient semantic retrieval.
- **To implement a Retrieval-Augmented Generation (RAG) framework** capable of generating factual, context-aware responses in Amharic.
- **To develop a user-friendly web interface** (using Flask) that enables seamless querying of the system in Amharic.
- **To ensure inclusivity and accessibility** by supporting natural Amharic queries and responses, bridging the gap between legal texts and everyday users.
- **To evaluate and fine-tune the system's performance** through accuracy testing, relevance assessment, and user experience feedback.
- **To contribute to Ethiopia's digital transformation efforts** by advancing the use of AI and NLP technologies for low-resource languages like Amharic.
- **To enhance legal literacy** among Ethiopian citizens by simplifying access to reliable legal resources.

- **To provide complementary tools,** such **as sentiment analysis,** to broaden the exploration of NLP applications in Amharic.

2.4 METHODOLOGY

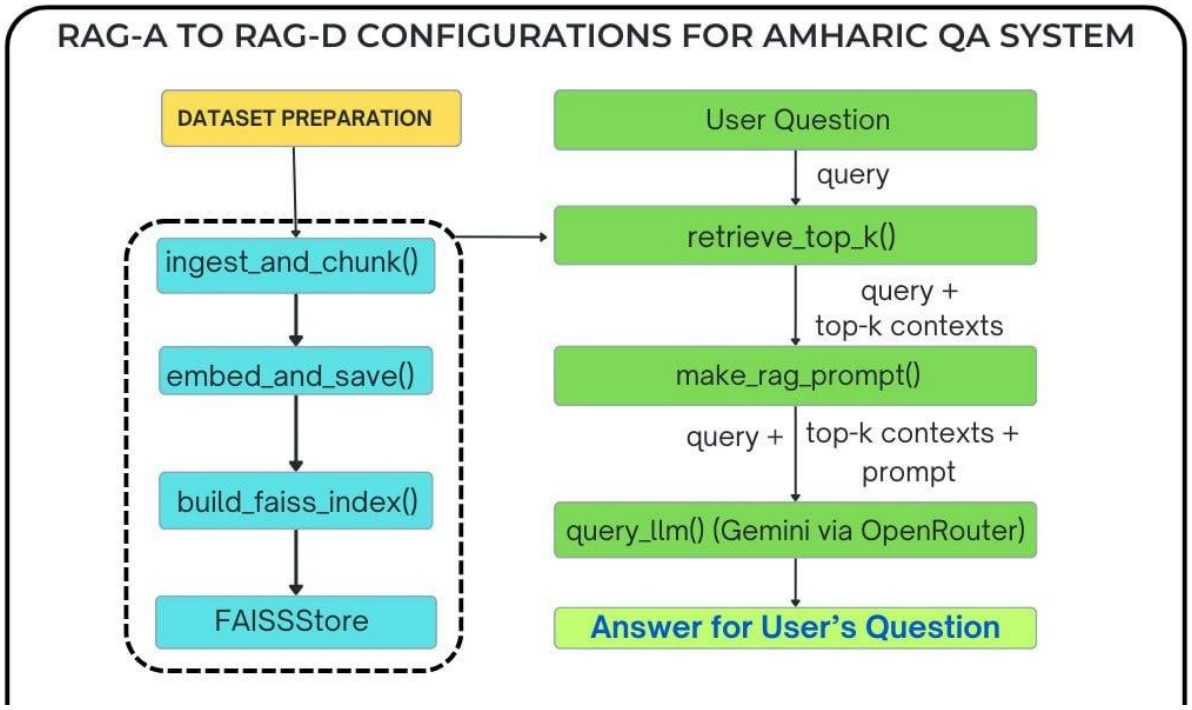
The methodology of this project adopted a systematic, step-by-step process to design, implement, and evaluate the Amharic Legal Chatbot. The workflow combined elements of data engineering, **natural language processing (NLP), machine learning (ML), and web development**, ensuring that the final system was both accurate and user-friendly, while remaining scalable for future enhancements.

Project development and testing were primarily carried out on Google Colab, leveraging its computational resources, while the deployment phase was implemented through a Flask-based web application to provide a practical and accessible interface for end users.

2.4.1 SYSTEM ARCHITECTURE OVERVIEW

The architecture of the **Amharic Legal Chatbot** is organized into **two primary phases**:

1. **Chatbot Development (Data Preparation and Knowledge Base Construction)**
2. **User Interface and Application Development (Flask-based Web App)**



Knowledge Base Construction (Data Preparation Phase)

The initial stage focused on preparing and organizing authentic Amharic legal texts to serve as the foundation of the chatbot's knowledge base.

- **Data Collection**: Legal documents, including **proclamations, regulations, and directives**, were collected from official Ethiopian government websites using **Python-based web scraping scripts**. Data was stored in **JSONL format**, with each entry including the extracted content and its source reference for **traceability**.
- **Normalization**: Raw Amharic texts were **cleaned and standardized** to remove formatting inconsistencies, spelling variations, and non-standard symbols.
- **Semantic Chunking**: To preserve context, the normalized text was divided into **semantic chunks of uniform size**, with each chunk overlapping the next by **two sentences** to ensure that critical legal details were not lost at chunk boundaries.
- **Embedding**: Each chunk was converted into **vector embeddings** using the **paraphrase-multilingual-mpnet-base-v2 model**, which is highly effective for **Amharic semantic similarity tasks**.
- **Vector Database Storage**: The embeddings, together with their associated text, were stored in a **Qdrant/FAISS vector database** for **efficient similarity-based retrieval**. The database was later exported as a **.pkl file** for integration into the chatbot system.or seamless integration into the chatbot system.

Application Development (Flask Web App)

After preparing the knowledge base, the chatbot was implemented using the **Flask framework**. The workflow includes:

- **User Query Submission – Users enter their legal questions in Amharic via the web-based interface.**
- **Similarity Search** – The application loads the **.pkl vector database** and identifies the top **N most semantically relevant text chunks** using **cosine similarity**.
- **Prompt** Assembly– The retrieved text segments are combined into a **structured, context-rich prompt** and sent to the **Gemini LLM**, with explicit instructions to rely exclusively on the retrieved content to **minimize hallucinated or incorrect responses**.

- **Response Generation** – If relevant matches are found, Gemini generates a **concise Amharic response**. If no suitable context exists, the system responds: *"I have no such information related to this question."*
- **Answer Delivery** – The generated response is displayed in a conversational flow on the web interface.

Flask Implementation

The Flask application served as the backbone of the chatbot, managing its core operational logic. Key functionalities of the system included:

Loading the precomputed **vector embeddings file**.

- Using **cosine similarity**, the system performs **efficient semantic searches** to identify the most relevant chunks of legal text for a given query.
- **Loading Precomputed Embeddings** – The application retrieves the vector embeddings previously generated and stored in the .pkl database.
- **Context-Enhanced Prompt Construction** - Retrieved text segments are integrated into **contextually enriched prompts** to guide the **Gemini LLM** in generating accurate and relevant responses.
- **Chat History Management** – The system maintains **session-based conversation history**, allowing for **multi-turn dialogues** and a more natural user interaction.

The entire chatbot workflow is encapsulated within the `index()` `route`, which orchestrates `query reception`, `context retrieval`, and `AI-driven response generation`, ensuring smooth and coherent interactions for the user.

2.4.2 DATA COLLECTION

The dataset comprised genuine Ethiopian legal documents, such as proclamations, regulations, and directives, predominantly written in Amharic. These documents were sourced from official government websites, guaranteeing credibility, accuracy, and reliability. Given that Ethiopian legal texts represent a low-resource domain, the dataset necessitated careful curation and meticulous preprocessing to ensure its suitability for AI-based applications.

```python
from googlesearch import search

def get_links_from_simple_search(query, num_results):
    """
    Retrieves URLs from a simple Google search.

    Args:
        query (str): The search query.
        num_results (int): The desired number of search results. This is not directly supported by googlesearch,
                           so it will be used to limit the results after fetching.

    Returns:
        list: A list of URLs.
    """
    # Fetch the search results
    search_results = search(query, num_results=num_results)
    # Convert the generator object to a list
    urls = list(set(search_results))
    return urls
```

2.4.3 DATA PREPROCESSING

To enhance the **quality of the dataset** and ensure it was suitable for embedding, the raw legal texts underwent a series of **systematic preprocessing steps**:

- **Text Cleaning** – Extraneous symbols, numbers, and formatting irregularities were **removed** to reduce noise and improve clarity.
- **Normalization** – Variations in **spelling and script** within Amharic texts were **standardized**, ensuring consistency across the dataset.
- **Tokenization** – Sentences were segmented into **individual words or sub-word units**, preparing the text for embedding models.
- **Chunking** – Lengthy documents were divided into **smaller, semantically coherent sections**, which enhanced **retrieval efficiency** during semantic searches.

These steps significantly improved **embedding accuracy** and **retrieval reliability**.

```python
import json

def read_jsonl_content(file_path, content_key='content'):
    """Returns a list of 'content' values from a .jsonl file."""
    content_values = []
    with open(file_path, 'r', encoding='utf-8') as file:
        for line in file:
            stripped_line = line.strip()
            if stripped_line:
                record = json.loads(stripped_line)
                content_values.append(record[content_key])

    print(f"Read {len(content_values)} records from {file_path}.")
    return content_values  # Example: ["text1", "text2", ...]
```

```python
def normalize_and_save_data(jsonl_data, file_path):
    """Writes all JSONL entries as one continuous text blob."""
    with open(file_path, 'w', encoding='utf-8') as f:
        all_text = " ".join(normalize_amharic(text) for text in jsonl_data)  # No newlines
        f.write(all_text)
```

```python
def normalize_jsonl_file(input_file, output_file):
  jsonl_data=read_jsonl_content(test_file_path)
  normalize_and_save_data(jsonl_data, destination_file_path)
```

```python
import os
import os
import shutil
from google.colab import files

def read_normalized_text(file_path):
    with open(file_path, 'r', encoding='utf-8') as f:
        return f.read()


def split_sentences(file_path: str) -> list:
    """
    Reads a .txt file and splits its content into sentences using ., !, or ? as delimiters.

    Args:
        file_path: Path to the input text file.

    Returns:
        List of sentences (strings), each ending with a delimiter.
    """
    with open(file_path, "r", encoding="utf-8") as f:
        text = f.read()
        text = re.sub(r'\(\s*\)', '', text)
    sentences = []
    current_sentence = ""

    for char in text:
        current_sentence += char
        if char in {'.', '!', '?'}:
            stripped = current_sentence.strip()
            if stripped:
                sentences.append(stripped)
            current_sentence = ""

    if current_sentence.strip():
        sentences.append(current_sentence.strip())

    return sentences


def chunk_and_save(sentences: list, output_dir: str = '/content/Scrapify/chunks',
                   max_chars: int = 1000, max_overlap_chars: int = 250):
    os.makedirs(output_dir, exist_ok=True)

    chunks = []
    current_chunk = ""
    current_sentences = []

    for sentence in sentences:
        if len(current_chunk) + len(sentence) <= max_chars:
            current_chunk += sentence
            current_sentences.append(sentence)
        else:
            # Determine overlap sentences
            overlap_sentences = []
            total_overlap = 0
```

```
            for s in reversed(current_sentences):
                if total_overlap + len(s) <= max_overlap_chars:
                    overlap_sentences.insert(0, s)
                    total_overlap += len(s)
                else:
                    break

            # Ensure at least one sentence is overlapped
            if not overlap_sentences and current_sentences:
                overlap_sentences = [current_sentences[-1]]

            chunks.append(current_chunk)

            # Start new chunk with the overlap + current sentence
            current_sentences = overlap_sentences + [sentence]
            current_chunk = ''.join(current_sentences)

    if current_chunk:
        chunks.append(current_chunk)

    # Save to files (uncomment if needed)
    # for i, chunk_data in enumerate(chunks):
    #     chunk_filename = f"{output_dir}/chunk_{i+1}.txt"
    #     with open(chunk_filename, 'w', encoding='utf-8') as f:
    #         f.write(chunk_data)

    return chunks

def zip_chunks_folder(folder_path):
    zip_file_path = f"{folder_path}.zip"
    shutil.make_archive(folder_path, 'zip', folder_path)
    return zip_file_path
```

```
sentences = [
    "Hello world::",
    "How are you today::",
    "This is a test sentence::",
    "Goodbye::",
    "How are you today1::",
    "How:How are you today:: w are you today::How are you today::How are youday:: y::How are you today::",
    "This is a test sentence::",
    "Goodbye::",
    "How are you today::",
    "This is a test sentence::",
    "Goodbye::",
    "How are you today::",
    "This is a test sentence::",
    "Goodbye::",
    "How are you today::",
    "This is a test sentence::",
    "Goodbye::"
]

chunks=chunk_and_save(sentences, max_chars=100, max_overlap_chars=25)
chunks[2],len(chunks[2])
```

2.4.4 EMBEDDING AND VECTORIZATION

Preprocessed The **preprocessed legal texts** were transformed into **semantic embeddings** using a **multilingual transformer model**, allowing the system to capture the **contextual meaning** of the content rather than relying solely on keyword matches. This enabled the chatbot to interpret and respond accurately to **Amharic-language queries**.

The resulting embeddings were **indexed and stored** in a **vector database** (such as **FAISS** or **Qdrant**). At query time, the system performed a **vector similarity search**

to retrieve the most relevant text segments, which then served as the **knowledge base grounding** for generating accurate and context-aware responses.

```python
# Import SentenceTransformer
from sentence_transformers import SentenceTransformer

# Load the multilingual SBERT model
model = SentenceTransformer('paraphrase-xlm-r-multilingual-v1')

# Upload and process the chunks
chunks = upload_and_unzip_chunks()

# Now generate embeddings for the chunks
if chunks:
    vectors = model.encode(chunks, show_progress_bar=True)
    print("Embeddings generated successfully.")
else:
    print("No chunks were processed.")
```

```python
import numpy as np

# Save embeddings to a file inside Colab
np.save('/tmp/embeddings.npy', vectors)
from google.colab import files

# Download the file to your PC
files.download('/tmp/embeddings.npy')
```

2.4.5 RETRIEVAL-AUGMENTED GENERATION (RAG) FRAMEWORK

The chatbot was implemented using a Retrieval-Augmented Generation (RAG) architecture, which integrates two main components:

1. **Retriever** – Conducts **semantic similarity searches** to identify and fetch the most relevant legal text segments from the knowledge base.

2. **Generator** – Utilizes the **Gemini LLM** to generate **coherent, context-aware responses in Amharic**, grounded in the retrieved documents.

This combination ensures that the chatbot produces answers that are both **factually accurate**—based on official legal texts and **naturally expressed**, enabling smooth, conversational interactions in **Amharic**.

2.4.6 WEB INTERFACE DEVELOPMENT

To maximize **accessibility and usability**, a **lightweight and responsive web application** was developed using the **Flask framework**. Key features of the interface include::

- **Amharic Query Input** – A dedicated input box allowing users to submit legal questions directly in **Amharic**.

- **Real-Time Contextual Responses** – The system delivers **instant, context-aware answers** based on the knowledge base.
- **User-Friendly Design** – A **clean and minimal interface** was implemented to accommodate users with **limited digital literacy**, ensuring smooth and intuitive interactions.

This design allowed the chatbot to provide **inclusive, efficient, and straightforward access** to legal information for a broad range of Ethiopian users.

2.4.7 EVALUATION AND TESTING

The system was assessed through a **two-tier evaluation approach**:

- **Technical Evaluation** – The chatbot's performance was measured by assessing **retrieval accuracy, relevance of responses, and computational efficiency** using a set of representative sample queries.
- **User Experience Testing:** Informal feedback was collected from **peers and mentors** to evaluate the system's **clarity, usability, and reliability**, ensuring that it met user expectations and provided an intuitive interaction experience.

2.4.8 TOOLS AND TECHNOLOGIES USED

The project leveraged a range of **programming tools, frameworks, and environments** to implement and deploy the chatbot:

- **Programming Language:** Python, HTML and CSS
- **Frameworks and Libraries:** Flask, Transformers (Hugging Face), Scikit-learn, Pandas, NumPy
- **Environment:** Google Colab, Jupyter Notebook, VS Code
- **Database:** Vector Database (FAISS) for semantic search
- **Version Control:** Git/GitHub for collaboration and backup

2.5 LITERATURE REVIEW

The The creation of the Amharic Legal Chatbot lies at the convergence of three key research areas: chatbot technologies, the application of Artificial Intelligence (AI) and Natural Language

Processing (NLP) in legal contexts, and processing of low-resource languages, with a particular emphasis on Amharic.

Examining prior studies in these domains provides a **contextual foundation**, identifies **existing research gaps**, and underscores the **novelty and significance** of this project. This review highlights how the current work builds upon earlier research while addressing unique challenges related to **Amharic-language legal information access**.

2.5.1 CHATBOTS: FROM RULE-BASED SYSTEMS TO AI-POWERED ASSISTANTS

Over the past several decades, chatbots have undergone substantial evolution. Early systems were rule-based, relying on predefined keywords and scripted responses, such as ELIZA in the 1960s. While these systems illustrated the potential of conversational agents, they were limited in flexibility and unable to handle queries beyond their programmed patterns.

With the advent of machine learning and NLP, modern chatbots shifted towards retrieval-based and generative architectures. Retrieval-based systems identify the most appropriate answer from an existing knowledge base, whereas generative systems employ deep learning models to dynamically generate responses. Advances in transformer-based architectures—including BERT, GPT, and LLaMA—have further enhanced chatbots' ability to comprehend context and provide coherent, human-like interactions.

Today, chatbots are widely used across customer service, healthcare, education, and government sectors, demonstrating their versatility in improving both accessibility and operational efficiency. Despite these global advancements, the deployment of chatbot technologies in Ethiopian legal contexts and the Amharic language remains limited, highlighting the need for localized solutions.

2.5.2 AI AND NLP FOR LEGAL TECHNOLOGY

The **legal domain** poses unique challenges for NLP due to the **length, formality, and complexity** of its texts, which often include **specialized vocabulary** and references to prior laws. Consequently, AI applications in legal technology have focused on tasks such as **document retrieval, text summarization, legal reasoning, and question answering**.**Legal Information Retrieval:** Research in legal AI has produced systems capable of efficiently retrieving case law, statutes, or regulations. Projects such as **COLIEE**

(Competition on Legal Information Extraction and Entailment) have advanced methods for legal document search and entailment. However, most of these systems are designed for English or other high-resource languages.

- **Legal Information Retrieval**: Research in legal AI has produced systems capable of efficiently **retrieving statutes, case law, and regulations**. Initiatives like **COLIEE (Competition on Legal Information Extraction and Entailment)** have advanced techniques for legal document search and entailment. However, most systems are designed for **English or other high-resource languages**, limiting their applicability to low-resource contexts.
- **Question Answering Systems**: Recent approaches leverage **Retrieval-Augmented Generation (RAG) frameworks** to answer legal queries by grounding generative models in structured legal data. Models such as **LegalBERT** and **CaseLaw-BERT** have demonstrated strong performance in retrieving relevant legal provisions and case law.
- **Challenges in Legal AI:** Despite these advances, legal NLP remains difficult due to several domain-specific issues:
    - **Ambiguity and context dependency** in legal language.
    - **Dynamic nature of law**, where regulations are frequently updated.
    - **High stakes of errors**, since inaccurate responses may lead to serious legal consequences.

In Ethiopia, access to legal information is already limited, and no large-scale legal AI systems exist in Amharic. This makes the development of an **Amharic Legal Chatbot** an important step toward bridging the accessibility gap by adapting global advances in **AI and legal NLP** to a **local, under-resourced context**.

2.5.3 NLP FOR LOW-RESOURCE LANGUAGES AND AMHARIC

While English, Chinese, and other global languages benefit from abundant datasets and pretrained models, **Amharic is a low-resource language** with limited annotated corpora, pretrained embeddings, and NLP tools. This scarcity presents challenges for developing robust AI applications.

- **Amharic-Specific Challenges:**

- **Morphological complexity** – Amharic is a Semitic language with rich inflection, making tokenization and normalization difficult.
- **Script limitations** – Written in the **Ge'ez script**, Amharic lacks standardized digital processing resources compared to Latin-based scripts.
- **Data scarcity** – Few large, clean, and publicly available Amharic corpora exist, limiting training opportunities for large models.

- **Research Efforts in Amharic NLP:** Despite these challenges, several promising initiatives have emerged:
  - Development of **Amharic part-of-speech taggers, morphological analyzers, and stemmers**.
  - Use of **multilingual transformer models** (e.g., mBERT, XLM-R, AfriBERTa) which support Amharic through cross-lingual transfer learning.
  - Early work in **Amharic sentiment analysis** and **machine translation**, showing feasibility of applying deep learning to low-resource languages.
- **Relevance to Legal Applications:** These developments provide a foundation for applying **semantic search and generative AI techniques** to Amharic legal documents. However, existing research is primarily academic, and **practical tools for public use remain limited**. The Amharic Legal Chatbot builds upon these early advances by demonstrating a **real-world deployment** of Amharic NLP in a **critical domain-legal accessibility**.

2.5.4 RESEARCH GAP AND CONTRIBUTION OF THE PROJECT

From the reviewed literature, it is evident that:

- Modern **chatbot architectures** are highly advanced but largely unavailable in Amharic.
- **Legal AI systems** exist globally but are concentrated on English or high-resource languages.
- Research in **Amharic NLP** has laid a foundation but has not yet been applied at scale to public-facing legal services.

The **Amharic Legal Chatbot** addresses these gaps by:

1. Adapting **RAG-based architectures** for the Amharic language.

2. Building a **vectorized legal knowledge base** from Ethiopian laws.
3. Providing a **user-facing, Amharic-language chatbot** that directly supports Ethiopian citizens, students, and professionals.

Thus, the project contributes both to the **global field of low-resource NLP** and to Ethiopia's national goals of **digital transformation and equitable access to justice**.

CHAPTER THREE

3. INTERNSHIP EXPERIENCE AND SPECIFIC WORK

3.1 SELECTION OF THE ORGANIZATION

When it came time to select an institution for my internship, I carefully chose the Information Network Security Administration (INSA). My decision was guided by several factors that aligned closely with my academic background as a fifth-year Software Engineering student at Arba Minch University (AMU) and my long-term ambitions in Artificial Intelligence and Software Development.

Firstly, INSA is among Ethiopia's most prestigious national institutions, tasked with safeguarding information infrastructure, advancing cybersecurity, and promoting strategic technologies such as AI and Natural Language Processing (NLP). As a software engineering student, I was eager to observe and engage with advanced AI techniques in a real-world, high-impact setting. INSA provided this unique opportunity, being one of the few organizations in Ethiopia actively investing in AI research and development.

Secondly, the internship placement aligned strongly with my academic interests. During my studies at AMU, I developed a keen passion for Machine Learning, NLP, and AI applications for low-resource languages, particularly Amharic. INSA's ongoing projects in Amharic language technologies, digital security, and knowledge-based systems offered an ideal environment to expand my expertise. Unlike private companies that may focus mainly on commercial software products, INSA combines research with national service, allowing interns to contribute both academically and practically.

Thirdly, the specific internship project—the development of an Amharic Legal Chatbot—perfectly aligned with my goal of creating impactful AI-driven solutions for Ethiopian society.

Working on a system designed to make legal information more accessible in Amharic resonated not only with my technical interests but also with my conviction that technology should serve social good. This made INSA the most appropriate and rewarding place to undertake my internship.

Finally, I was drawn to INSA because of its reputation as a structured, learning-oriented institution. I anticipated—and later experienced—that working under the guidance of expert mentors would expose me to professional workflows, collaborative teamwork, and the discipline required for large-scale projects. This environment served as a crucial bridge between academic theory and professional practice.In summary, I selected INSA because it provided:

- A **unique opportunity** to contribute to national AI and NLP projects, especially in Amharic.
- **Exposure to real-world applications** of AI in cybersecurity and information management.
- Strong **alignment between my academic background, research interests, and career aspirations**.
- A **structured, professional environment** with mentorship from experienced practitioners.

This choice proved to be **one of the most rewarding academic decisions** I have made, as the internship not only facilitated **technical growth** but also gave me a profound sense of contributing to **Ethiopia's digital transformation goals**.

3.2 PLACEMENT WITHIN THE ORGANIZATION

During my internship at the **Information Network Security Administration (INSA)**, I was assigned to the **Technology and Innovation Division**, specifically within the **Computing and Data Analytics Department**, under the **Data Analytics (AI & Machine Learning) Unit**. This placement closely aligned with both my **academic background** and the focus of my internship project, which involved developing an **Amharic Legal Chatbot** using advanced AI techniques.

The **Technology and Innovation Division** plays a pivotal role within INSA, spearheading initiatives that leverage **emerging technologies** to advance **national development and cybersecurity objectives**. Within this division, the **Computing and Data Analytics**

Department focuses on creating **intelligent systems** and **data-driven solutions** to strengthen Ethiopia's technological autonomy. The **AI & Machine Learning Unit**, where I was placed, specializes in **advanced analytics, natural language processing, and machine learning applications** aimed at addressing critical national challenges.

My placement in this unit was strategically advantageous for several reasons:

1. **Alignment with Project Goals** – Since my project involved developing an **AI-powered chatbot for Amharic legal texts**, being part of the AI & ML team provided direct access to **technical expertise, computational resources**, and collaborative discussions on **NLP workflows and machine learning methodologies**.
2. **Exposure to Advanced Research and Applications** – The unit actively engages in projects spanning **natural language understanding, sentiment analysis, and predictive analytics**, broadening my perspective on how AI can be applied not only to **legal chatbots** but also to **cybersecurity, governance, and data intelligence initiatives**.
3. **Hands-On Mentorship and Collaboration** – Placement in the AI & ML unit allowed me to work under the mentorship of experienced professionals, including **Mr. Terefe Feyisa**, who provided guidance on both **technical and project management aspects**. I also collaborated with **data scientists, researchers, and software engineers**, gaining practical experience in **teamwork and interdisciplinary collaboration**.
4. **Strategic Significance of the Unit** – The placement highlighted INSA's recognition of **AI as a strategic driver** for national security and technological advancement. By contributing to this unit, I gained **practical experience** while also participating in a **national initiative** to advance **Amharic language technologies**.

In summary, my assignment within the **Technology and Innovation Division (Computing and Data Analytics Department, AI & Machine Learning Unit)** was exceptionally suitable. It provided a **supportive and resource-rich environment** to design, implement, and evaluate my internship project, while also offering valuable insight into how **AI and ML are integrated into Ethiopia's broader technological and cybersecurity strategies**.

3.3 WORKFLOW IN THE ASSIGNED SECTION

The workflow within the AI & Machine Learning Unit of the Computing and Data Analytics Department was highly structured and research-oriented, following a research-and-

development (R&D) model. Since the unit primarily focused on addressing complex technological challenges, its workflow was organized into sequential phases including data engineering, model development, system integration, and evaluation cycles.

3.3.1 PLANNING AND TASK ASSIGNMENT

At the start of each project cycle, the team conducted planning meetings led by a team leader or senior data scientist. These sessions were used to define project objectives, assign responsibilities, and break down larger initiatives into smaller, manageable tasks.

For instance, in my **Amharic Legal Chatbot project**, the tasks were organized into:

- Data Collection and Preprocessing
- Embedding generation and vectorization
- Model training and fine-tuning
- Web application integration

This ensured clear accountability and collaborative progress.

3.3.2 DATA ENGINEERING WORKFLOW

Data played a central role in most of the unit's projects. The team followed a structured pipeline:

- **Data Acquisition** – Collecting raw data from trusted sources such as government portals, research datasets, and local repositories.
- **Data Cleaning & Normalization** – Removing duplicates, correcting inconsistencies, and standardizing Amharic text.
- **Annotation & Structuring** – When needed, the team annotated datasets for supervised learning tasks such as sentiment analysis or classification.
- **Database Management** – Storing structured datasets in **vector databases (e.g., FAISS/Qdrant)** for efficient semantic search.

This workflow ensured that downstream machine learning tasks had **high-quality, usable data**.

3.3.3 MODEL DEVELOPMENT WORKFLOW

The core of the unit's operations was **AI model development and experimentation**. This involved:

- **Choosing suitable architectures** (transformer-based models such as multilingual BERT or mpnet).
- **Generating embeddings** for semantic similarity in Amharic.
- **Training/fine-tuning models** for specific applications (e.g., legal chatbot, sentiment analysis).
- **Evaluating models** using accuracy, F1-score, and relevance-based metrics.

Model development often followed an **iterative process**: experiments were run in **Google Colab or local servers**, results were reviewed, and parameters were adjusted to improve performance.

3.3.4 INTEGRATION AND APPLICATION DEVELOPMENT

Once models were trained, they were integrated into real-world systems. In my case, this involved:

- Exporting vector databases to **.pkl format**.
- Building a **Flask web application** that handled user input, semantic retrieval, and response generation.
- Connecting the system to an **LLM (Gemini)** for natural language generation.
- Designing a **simple front-end** for ease of access.

This stage highlighted the **collaboration between data scientists and software engineers**, as integration required both AI knowledge and software development skills.

3.3.5 REVIEW AND TESTING

The team placed strong emphasis on **testing and validation**. Workflow here included:

- **Unit Testing** – Checking if each pipeline step (data preprocessing, embedding search, response generation) worked correctly.
- **System Testing** – Running full chatbot sessions to ensure smooth interaction.
- **Peer Review** – Presenting progress to senior experts for feedback.

- **User Testing** – Collecting informal feedback from peers and mentors.

3.3.6 DOCUMENTATION AND KNOWLEDGE SHARING

Finally, every workflow cycle was concluded with proper **documentation and reporting**. The team maintained:

- **Code Repositories (GitHub)** for version control.
- **Technical Documentation** explaining models, datasets, and system workflows.
- **Knowledge Sharing Sessions** where team members presented their work, challenges, and lessons learned.

Generally, the **workflow of the AI & Machine Learning Unit** combined elements of academic research, industry-standard engineering practices, and national-level problem solving. It ensured that projects like the **Amharic Legal Chatbot** were not only technically sound but also strategically aligned with INSA's mission of advancing indigenous technology and digital security.

3.4 SPECIFIC WORK TASKS EXECUTED

During my internship at **INSA, Technology and Innovation Division, Computing and Data Analytics Department (AI & Machine Learning Unit)**, I was actively engaged in both **research-oriented** and **application-driven** tasks. My work revolved around the **design, development, and deployment of the Amharic Legal Chatbot**, with supplementary exposure to other AI workflows within the unit. Below are the major tasks I executed:

3.4.1 DATA COLLECTION AND PREPARATION

- I **scraped Ethiopian legal documents** (proclamations, regulations, and directives) from trusted government sources.
- Extracted texts were structured into **JSONL files**, including both the content and source references.
- Conducted **text cleaning and normalization** to remove redundant symbols, inconsistencies, and non-standard Amharic scripts.
- Implemented **semantic chunking** of documents into smaller passages with overlapping sentences to ensure **context preservation**.

This was one of the most critical steps, as the **quality of input data** directly impacted the chatbot's performance.

3.4.2 EMBEDDING AND VECTOR DATABASE DEVELOPMENT

- Converted Amharic legal text chunks into **dense embeddings** using the **paraphrase-multilingual-mpnet-base-v2** model.
- Stored embeddings along with their original texts in a **Qdrant vector database**.
- Exported the final database into a **.pkl file** for efficient integration with the chatbot.

This allowed the system to perform **semantic similarity search**, ensuring that user queries retrieved the most relevant sections of Ethiopian law.

3.4.3 CHATBOT INTEGRATION WITH RETRIEVAL-AUGMENTED GENERATION (RAG)

- Designed the pipeline where user queries were first matched with legal text chunks from the **vector database**.
- Constructed structured **prompts** combining retrieved legal texts and user questions.
- Integrated the prompts with **Gemini LLM**, ensuring that generated responses were **fact-based** and **contextually grounded**.
- Implemented fallback behavior where the system politely responded:

"I have no such information related to this question." if no relevant data was retrieved.

This ensured reliability and minimized **AI hallucinations**.

3.4.4 FLASK WEB APPLICATION DEVELOPMENT

- Developed a **Flask-based web interface** to make the chatbot accessible.
- Implemented routes for:
    - **User input submission**
    - **Embedding similarity search**
    - **Response generation and delivery**
- Added a **chat history session** to allow continuous, conversational interaction.

- Focused on building a **simple, intuitive, and responsive UI** so that ordinary citizens could easily use the system.

3.4.5 SUPPLEMENTARY SENTIMENT ANALYSIS PROJECT

- As part of skill enhancement, I worked on an **Amharic Sentiment Analysis model**.
- Preprocessed Amharic text datasets and trained ML models to classify text into **positive, negative, or neutral** categories.
- Gained hands-on experience in **feature extraction, training pipelines, and model evaluation**.

Although secondary to the chatbot, this project deepened my understanding of **Amharic NLP challenges**.

3.4.6 TESTING, EVALUATION, AND DOCUMENTATION

- Conducted **technical evaluation** of the chatbot: checking retrieval accuracy, runtime performance, and relevance of responses.
- Performed **sample user testing** with peers and mentors to collect feedback on usability.
- Debugged issues related to **encoding errors, mismatched responses, and UI rendering**.
- Documented all processes including **data pipelines, code workflows, and system architecture** for future maintainability.

3.5 METHODS, TOOLS, AND TECHNIQUES USED

During my internship at **INSA**, I applied a wide range of methods, tools, and techniques drawn from **AI, Natural Language Processing (NLP), Machine Learning (ML), and Software Engineering**. These methodologies were essential in ensuring that the **Amharic Legal Chatbot** was both technically sound and user-friendly. Below are the main categories:

3.5.1 DATA ENGINEERING & PREPROCESSING METHODS

Since Ethiopian legal documents are **lengthy, unstructured, and complex**, preprocessing was crucial. I used:

- **Web Scraping Techniques** → Python scripts (BeautifulSoup, Requests) to extract proclamations and directives from government websites.

- **Text Cleaning** → Removed special characters, redundant symbols, and irrelevant formatting.
- **Normalization** → Standardized Amharic Unicode representations and corrected spelling inconsistencies.
- **Tokenization** → Split sentences into words and subwords using NLP libraries (e.g., Hugging Face tokenizers).
- **Semantic Chunking** → Broke large legal texts into smaller overlapping passages to improve retrieval accuracy.

Impact: Improved the **consistency and quality** of legal data, making it machine-readable for AI models.

3.5.2 AI/NLP AND MACHINE LEARNING METHODS

To build a chatbot that could understand and respond in Amharic, I applied several AI techniques:

- **Vector Embeddings**
  - Used **paraphrase-multilingual-mpnet-base-v2** transformer model for generating semantic embeddings.
  - Captured the **meaning of Amharic text**, enabling the chatbot to retrieve relevant information beyond simple keyword matching.
- **Semantic Similarity Search**
  - Stored embeddings in a **Qdrant Vector Database** and later in a .pkl format for fast retrieval.
  - Applied **cosine similarity** to rank legal text chunks most relevant to the user's query.
- **Retrieval-Augmented Generation (RAG)**
  - Combined **retriever** (semantic search) and **generator** (Gemini LLM) components.
  - Ensured responses were **fact-based**, avoiding unsupported AI hallucinations.
- **Sentiment Analysis (Side Project)**

- Implemented a text classification pipeline in Amharic.
- Applied **feature extraction**, **ML models** (e.g., Logistic Regression, Naïve Bayes), and **evaluation metrics** (accuracy, F1-score).

Impact: Enabled the chatbot to provide **reliable, context-grounded Amharic answers** while also showcasing the potential of Amharic NLP models.

3.5.3 SOFTWARE DEVELOPMENT TOOLS & FRAMEWORKS

To integrate AI models into a working system, I relied on:

- **Flask (Python Web Framework)** → Built the chatbot's backend, handling query processing, embedding retrieval, and LLM integration.
- **HTML & CSS** → Designed the front-end interface, ensuring user-friendliness.
- **Google Colab / Jupyter Notebook** → For model experimentation, data preprocessing, and debugging.
- **VS Code** → For application development and final deployment.
- **Git & GitHub** → Version control, code backup, and collaborative work.

Impact: Helped me transition from **research experiments** to a **deployable real-world web application**.

3.5.4 DATABASES AND STORAGE

- **Qdrant Vector Database** → Stored embeddings for semantic similarity search.
- **Pickle (.pkl) Files** → Exported embeddings for fast integration with the chatbot.
- **JSONL Files** → Stored structured legal texts with metadata.

Impact: Provided an efficient and scalable way to handle large volumes of Ethiopian legal texts.

3.5.5 EVALUATION & TESTING TECHNIQUES

- **Technical Testing** → Checked retrieval accuracy, semantic similarity relevance, and LLM response quality.
- **Error Analysis** → Identified cases where queries failed or responses were mismatched.
- **User Feedback** → Collected informal feedback from mentors and peers on chatbot usability.

- **Benchmarking** → Compared performance with basic keyword-based search to demonstrate the advantage of semantic search.

Impact: Improved chatbot reliability and highlighted the strengths of using AI for legal tech.

3.5.6 MECHANICAL ENGINEERING RELEVANCE (ACADEMIC INTEGRATION)

Even though the project was AI-focused, I applied **engineering methods** such as:

- **Problem Definition & Justification** → Identifying gaps in Amharic legal access.
- **Systematic Design Process** → Iterative pipeline from data collection → preprocessing → AI modeling → web deployment.
- **Optimization Techniques** → Ensured faster retrieval by testing chunk size and embedding dimensions.
- **Validation & Verification** → Ensured outputs matched legal sources and user needs.

Impact: Reinforced my **engineering mindset** structured problem-solving, optimization, and validation applied in a new digital context.

3.6 MAJOR CHALLENGES AND PROBLEMS FACED

During the course of my internship at INSA, I faced several challenges that tested both my technical skills and problem-solving abilities. These difficulties arose from the nature of the project itself, the tools and resources available, and the unique characteristics of working with Amharic language data. The major challenges are outlined below:

3.6.1 SCARCITY OF AMHARIC NLP RESOURCES

One of the most pressing challenges was the lack of pre-built natural language processing tools for Amharic. Unlike English or other widely spoken languages, Amharic has limited resources such as annotated datasets, pretrained embeddings, or standardized preprocessing pipelines. As a result, many tasks such as tokenization, normalization, and semantic chunking had to be carefully customized.

3.6.2 DATA QUALITY AND STRUCTURE ISSUES

Legal documents collected from Ethiopian government portals were often lengthy, unstructured, and inconsistent in formatting. Some texts contained errors, overlapping sections,

or redundant metadata. Cleaning this data was time-intensive and required several iterations of preprocessing to ensure quality.

3.6.3 COMPUTATIONAL CONSTRAINTS

Although Google Colab provided a convenient platform for experimentation, its limited runtime sessions, restricted GPU usage, and memory constraints made large-scale model training and fine-tuning challenging. Handling long Amharic documents and generating embeddings for thousands of text chunks strained the resources.

3.6.4 INTEGRATION OF AI MODELS INTO A WEB APPLICATION

While model development and testing in Colab were manageable, deploying these models into a functional web interface required bridging research-level scripts with production-ready code. Adapting Python notebooks into Flask applications involved careful debugging, handling memory usage, and ensuring that the embedding database could be efficiently queried.

3.6.5 RISK OF AI HALLUCINATIONS

Large Language Models (LLMs), including Gemini, sometimes generated outputs not grounded in the retrieved legal context. This posed a significant risk, as incorrect legal information could mislead users. Designing effective prompts and ensuring strict reliance on retrieved contexts was a recurring challenge.

3.6.6 LIMITED COLLABORATION AND TESTING FEEDBACK

Since the project was highly technical and language-specific, only a limited pool of colleagues and mentors could provide meaningful feedback. This made the iterative testing and validation process slower, as fewer users were available to test the chatbot under real-world scenarios.

3.6.7 TIME CONSTRAINTS

The internship duration, while long enough for substantial learning, was still limited in the context of developing a full-fledged AI application. Balancing structured learning, experimentation, data preparation, model building, and deployment within the given timeframe was challenging.

3.7 MEASURES TAKEN / PROPOSED SOLUTIONS

To address the challenges encountered during the internship, I applied a combination of technical strategies, problem-solving approaches, and practical workarounds. These measures not only helped in overcoming obstacles but also strengthened my ability to deliver a functioning and reliable chatbot system.

3.7.1 TACKLING THE SCARCITY OF AMHARIC NLP RESOURCES

Since Amharic resources were limited, I relied on multilingual transformer models (such as *paraphrase-multilingual-mpnet-base-v2*) that included partial support for Amharic. I also designed **custom preprocessing pipelines** tailored to the language, which involved:

- Creating my own tokenization and normalization scripts.
- Applying semantic chunking with overlapping sections to preserve context.
- Conducting small-scale manual validation to verify preprocessing quality.

This ensured that Amharic texts could still be meaningfully processed despite the lack of specialized tools.

3.7.2 IMPROVING DATA QUALITY AND STRUCTURE

I developed a **data cleaning pipeline** to address unstructured legal texts. This pipeline involved:

- Removing duplicate sections, irrelevant formatting, and redundant metadata.
- Standardizing Amharic scripts to handle variations in spelling and Unicode usage.
- Organizing the cleaned documents into JSONL format, which stored both text and metadata such as the legal source.

This method improved data reliability, making the documents ready for embedding and retrieval.

3.7.3 OVERCOMING COMPUTATIONAL CONSTRAINTS

To deal with the limitations of Google Colab:

- I optimized the size of text chunks before embedding to balance memory use and retrieval accuracy.
- Used smaller batch sizes for embedding generation.
- Exported embeddings into .pkl files, allowing efficient reuse without re-running the embedding process every session.
- Where possible, I shifted lightweight tasks to local execution and reserved Colab GPUs for more demanding computations.

This hybrid approach made the workflow more efficient under resource constraints.

3.7.4 ENSURING SMOOTH INTEGRATION WITH FLASK

The transition from Colab notebooks to a web-based application required structural changes in the codebase. To manage this:

- I modularized the code, separating data retrieval, similarity search, and response generation into distinct functions.
- Tested each function independently before integrating into Flask routes.
- Used session management to maintain chat history, providing a natural conversational flow.

These steps improved system stability and usability.

3.7.5 REDUCING AI HALLUCINATIONS

To minimize the risk of generating unsupported legal advice:

- I implemented a **strict RAG framework** where the LLM was instructed to only answer from retrieved contexts.
- Added a safeguard response: *"I have no such information related to this question."* when no reliable match was found.
- Regularly tested the system with diverse queries to check if the responses were aligned with actual legal content.

This improved user trust and response credibility.

3.7.6 EXPANDING TESTING AND FEEDBACK

Although peer testers were limited, I encouraged my dormmates, colleagues, and mentors to interact with the chatbot. Their feedback on usability, clarity, and errors was valuable. I also created **sample test queries** representing common legal issues to evaluate chatbot consistency.

3.7.7 MANAGING TIME CONSTRAINTS

To balance the project workload, I adopted a structured schedule:

- Allocated the first weeks to data collection and preprocessing.
- Focused mid-phase on embedding generation, database creation, and model testing.
- Reserved the final weeks for web deployment, testing, and documentation.

This time management strategy allowed me to complete a working prototype within the internship period.

3.8 RESULTS AND DISCUSSION

The internship project produced a functional prototype of the **Amharic Legal Chatbot**, which successfully demonstrated the feasibility of applying advanced AI and NLP methods to Ethiopian legal texts. The outcomes can be grouped into two main categories: **technical achievements** and **personal/professional development**.

3.8.1 TECHNICAL ACHIEVEMENTS

3.8.1.1 Functional Chatbot System

- A working prototype of the Amharic Legal Chatbot was developed using a **Retrieval-Augmented Generation (RAG)** framework.
- The system allows users to enter queries in Amharic and receive **fact-based responses** retrieved from authentic legal documents.
- When queries fall outside the knowledge base, the chatbot provides a fallback response, ensuring **transparency and reliability**.
- The chatbot was deployed as a **Flask-based web application**, accessible via a user-friendly interface designed with simplicity in mind.

3.8.1.2 Data Pipeline and Knowledge Base

- A **structured dataset** of Ethiopian proclamations, directives, and regulations was collected, cleaned, and stored.
- The dataset was transformed into embeddings using **multilingual transformer models**, capturing semantic meaning in Amharic.
- A **vector database (Qdrant / Pickle format)** was implemented to support **fast similarity search**.

3.8.1.3 Integration of AI Models

- Successfully integrated the embedding-based retriever with **Gemini LLM** for context-aware response generation.
- Developed strict prompt engineering methods to ensure the LLM did not “hallucinate” or generate content beyond retrieved contexts.

- Demonstrated how **semantic similarity search** outperforms traditional keyword-based search for complex legal queries.

3.8.1.4 Sentiment Analysis Sub-Project

- Built a smaller side project for **Amharic sentiment classification** (positive, negative, neutral).
- Used feature extraction and ML models such as Logistic Regression and Naïve Bayes.
- Though not directly tied to the legal chatbot, this sub-project deepened my understanding of **text classification techniques** in low-resource languages.

3.8.2 PROFESSIONAL AND ACADEMIC LEARNING

3.8.2.1 Technical Growth

- Strengthened programming skills in **Python, Flask, and NLP libraries**.
- Gained hands-on experience in **data preprocessing, embedding generation, semantic search, and AI integration**.
- Learned how to bridge the gap between **research prototypes** and **practical, user-facing applications**.

3.8.2.2 Soft Skills Development

- Improved **time management** by structuring the internship phases effectively.
- Enhanced **problem-solving abilities** by overcoming challenges such as Amharic resource scarcity and computational constraints.
- Practiced **teamwork and communication**, collaborating with mentors and peers at INSA for feedback and guidance.

3.8.2.3 Societal Impact

- The project demonstrated how AI can address **real-world challenges in Ethiopia**, especially for under-resourced languages like Amharic.
- The chatbot has potential to improve **legal literacy**, reduce barriers to information, and support Ethiopia's **Digital Transformation goals**.
- The work highlighted the broader importance of **ethical and context-aware AI applications** in sensitive domains such as law.

3.8.3 DISCUSSION

The project highlighted both the **promise and challenges** of building AI applications in the Ethiopian context:

- On the positive side, the chatbot successfully showed that **AI can make legal information more accessible and understandable**, especially when grounded in authentic documents.
- On the other hand, challenges such as **data scarcity, limited computing resources, and risk of AI hallucinations** remain obstacles for scaling such projects.
- Despite these limitations, the project provides a **foundation for future research and development**, with potential to expand the system into other Ethiopian languages and domains.

3.9 RECOMMENDATIONS

Based on my internship experience and the outcomes of the **Amharic Legal Chatbot project**, several recommendations can be made to enhance the **performance, scalability, and sustainability** of similar initiatives. These suggestions target both **technical development** and **institutional support mechanisms**.

3.9.1 EXPAND AMHARIC NLP RESOURCES

One of the main bottlenecks was the scarcity of Amharic NLP tools and datasets. To overcome this, I recommend:

- Developing open-source Amharic tokenizers, stemmers, and lemmatizers.
- Building **large-scale annotated datasets** for Amharic, covering domains such as law, health, and education.
- Encouraging **research collaborations** between universities, government institutions, and tech companies to accelerate resource development.

3.9.2 STRENGTHEN DATA INFRASTRUCTURE

High-quality, structured data is the foundation of AI systems. Future efforts should:

- Establish a **centralized digital repository** of Ethiopian legal documents in machine-readable formats (JSON, XML).

- Ensure that documents are **regularly updated** as new proclamations and regulations are published.
- Introduce **metadata tagging standards** (such as categories, dates, and references) to improve searchability and traceability.

3.9.3 ENHANCE COMPUTATIONAL RESOURCES

The project was constrained by limited GPU access and storage capacity. For more robust systems:

- Provide **dedicated servers or cloud-based platforms** for NLP and AI development.
- Allocate **high-performance GPUs/TPUs** for research institutions like INSA and universities.
- Adopt **scalable infrastructure** (e.g., Docker containers, Kubernetes) for deployment and testing.

3.9.4 IMPROVE MODEL RELIABILITY AND SAFETY

Since legal information is sensitive, accuracy and trustworthiness must be prioritized. Suggested measures include:

- Strengthening **retrieval accuracy** by experimenting with specialized Amharic embeddings or fine-tuned transformer models.
- Expanding **prompt engineering** techniques to further minimize AI hallucinations.
- Implementing a **human-in-the-loop system**, where legal experts periodically review chatbot outputs for correctness.

3.9.5 EXPAND ACCESS AND USABILITY

To ensure wider adoption by citizens, the chatbot should be:

- Deployed as a **mobile-friendly app** (Android/iOS), since smartphones are the most common access point for Ethiopians.
- Enhanced with **voice-to-text** and **text-to-speech** features for accessibility, especially for users with low literacy.
- Designed with **multi-language support**, extending the chatbot to other Ethiopian languages such as Afaan Oromo, Tigrinya, and Somali.

3.9.6 INSTITUTIONAL AND POLICY SUPPORT

To maximize societal impact:

- INSA and similar institutions should **formally integrate digital legal assistants** into their e-government services.
- Encourage **policy-level support** for digitization of legal information in line with *Digital Ethiopia 2025*.
- Promote **legal awareness campaigns** that educate citizens about the availability and use of such digital tools.

3.9.7 FUTURE RESEARCH DIRECTIONS

Finally, for academic and technical advancement:

- Explore **domain adaptation techniques** to fine-tune large language models specifically for Ethiopian law.
- Investigate **knowledge graph integration** for structured legal reasoning.
- Research the use of **low-resource machine translation** to make Ethiopian legal documents available in multiple languages.

CONCLUSION OF RECOMMENDATIONS

Implementing these measures would significantly enhance the effectiveness, reliability, and scalability of the Amharic Legal Chatbot. More importantly, it would contribute to a broader vision of leveraging AI for low-resource languages, bridging the digital divide, and promoting equitable access to justice and legal information for Ethiopian citizens.

CHAPTER FOUR

4. BENEFITS GAINED FROM THE INTERNSHIP

My internship at the Information Network Security Administration (INSA), within the Technology and Innovation Division – Data Analytics (AI & Machine Learning), was a transformative learning experience. It allowed me to bridge the gap between academic knowledge and real-world applications while also cultivating professional, interpersonal, and life skills essential for future career success.

4.1 PRACTICAL SKILLS IMPROVEMENT

A major benefit of the internship was the improvement of my hands-on technical expertise. Through active participation in the *Amharic Legal Chatbot* project, I gained practical experience in:

- Preprocessing, cleaning, and normalizing Amharic legal documents.
- Implementing advanced NLP techniques such as embedding generation, semantic similarity search, and Retrieval-Augmented Generation (RAG).
- Utilizing modern tools including Flask, Hugging Face Transformers, Qdrant/FAISS databases, and Google Colab for model building and deployment.
- Developing a functional web application interface using Python, HTML, and CSS.

This practical engagement enabled me to translate theoretical concepts into working solutions, thereby strengthening both my programming and problem-solving abilities.

4.2 THEORETICAL KNOWLEDGE UPGRADING

In addition to practice, the internship deepened my theoretical understanding in several domains:

- Core principles of Natural Language Processing (NLP), particularly for low-resource languages such as Amharic.
- Transformer-based architectures and the semantic representation of embeddings.
- Legal informatics, focusing on structuring, indexing, and retrieving legal documents.

- Broader exposure to cybersecurity and digital governance frameworks within INSA's mandate.

This combination of theory and application enriched my academic perspective and laid the groundwork for future research and advanced studies.

4.3 INDUSTRIAL PROBLEM-SOLVING CAPABILITY

A unique benefit of working at INSA was the opportunity to engage with **real-world industrial challenges**. Unlike classroom projects, the problems faced in this environment were **open-ended, complex, and resource-constrained**. For example:

- Handling **incomplete or inconsistent legal datasets** required creative approaches to normalization and chunking.
- Developing **accurate yet efficient models** meant balancing computational resources with user needs.
- Ensuring **trustworthy outputs** in a sensitive domain like law required designing safeguards against hallucinations.

These experiences trained me to think critically, evaluate trade-offs, and design **practical solutions** rather than purely theoretical ones.

4.4 TEAMWORK AND COLLABORATION SKILLS

The internship emphasized the value of teamwork. I collaborated with mentors, peers, and domain experts in law and technology. Through this, I learned:

- How to **communicate technical ideas clearly** to non-technical team members.
- The importance of **knowledge sharing**, code reviews, and version control using GitHub.
- How to adapt to different working styles and contribute meaningfully in a **multi-disciplinary team**.

This strengthened my ability to function effectively in collaborative environments, an essential skill in both industry and research.

4.5 LEADERSHIP SKILLS DEVELOPMENT

Despite being an intern, I was entrusted with leadership responsibilities in certain components of the **Amharic Legal Chatbot** project, particularly in data preprocessing and embedding generation. This role allowed me to practice:

- Taking initiative and leading problem-solving efforts.
- Facilitating discussions around technical design choices.
- Managing tasks independently while ensuring coordination with teammates.

These responsibilities enhanced my confidence and prepared me for future leadership roles.

4.6 UNDERSTANDING OF WORK ETHICS AND INDUSTRIAL PSYCHOLOGY

The internship provided firsthand exposure to workplace discipline and ethics. I learned the importance of:

- Upholding confidentiality and integrity when handling sensitive national data.
- Practicing accountability and time management to meet deadlines consistently.
- Appreciating elements of industrial psychology, including motivation, stress management, and fostering positive professional relationships.

Such values are as critical as technical expertise for long-term career success.

4.7 ENTREPRENEURSHIP SKILLS

By working on a real-world problem with social impact, I gained insights into **entrepreneurial thinking**. Developing the Amharic Legal Chatbot showed me:

- How innovative ideas can address **local community needs**.
- The potential of AI solutions to create **startup opportunities**, particularly in low-resource language technology.
- The process of moving from **concept to prototype to deployment**, which mirrors the innovation lifecycle in entrepreneurship.

This motivated me to think beyond academic research and consider **AI-driven ventures** as a pathway for future career development.

4.8 INTERPERSONAL COMMUNICATION SKILLS

Finally, the internship significantly enhanced my interpersonal and communication skills. Through regular interaction with supervisors, peers, and domain experts, I learned to:

- Present complex technical ideas in clear and accessible terms.
- Listen actively to feedback and integrate it constructively.
- Adjust communication styles when engaging with technical professionals, legal specialists, or non-technical users.

These skills boosted my confidence in engaging diverse audiences, an essential ability for both professional and social interactions.

CONCLUSION OF BENEFITS GAINED

Overall, this internship was a transformative experience that combined technical mastery with personal and professional growth. It enhanced my expertise in AI, NLP, and low-resource language technologies while also strengthening my abilities in problem-solving, teamwork, leadership, ethics, entrepreneurship, and communication. Together, these skills have prepared me to be a more capable engineer, researcher, and innovator in the field of Artificial Intelligence and beyond.

CHAPTER FIVE

5. CONCLUSION AND RECOMMENDATIONS

5.1 GENERAL CONCLUSION

My internship at the Information Network Security Administration (INSA), held from February 8 to June 15, 2025, has been an invaluable and transformative experience. It allowed me to engage deeply with both the practical and theoretical aspects of Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP), with a particular emphasis on Amharic language technologies.

The central focus of my work was the development of an **Amharic Legal Chatbot**, designed to improve access to legal information for Ethiopian citizens. In a country where Amharic is widely spoken but underrepresented in digital technologies, this project addressed not only a technical challenge but also a socially significant issue. By implementing a Retrieval-Augmented Generation (RAG) framework, the chatbot demonstrated how AI can bridge the gap between citizens and essential legal knowledge.

Through this project, I was able to integrate data engineering, NLP, system design, and web development into a fully functional prototype aligned with Ethiopia's **Digital Transformation Strategy (Digital Ethiopia 2025)**. I also gained complementary experience in sentiment analysis for Amharic text, which expanded my knowledge of text classification methods.

Beyond technical expertise, the internship helped me cultivate essential soft skills, including teamwork, leadership, problem-solving, communication, and professional ethics. Working within a national institution highlighted the importance of confidentiality, accountability, and resilience when handling sensitive and mission-critical data.

In summary, this internship enabled me to:

- Apply classroom knowledge to real-world projects with tangible societal impact.
- Strengthen my technical expertise in advanced AI and NLP tools.
- Develop problem-solving strategies in complex, resource-constrained environments.

- Enhance my teamwork and leadership abilities while learning workplace discipline.
- Contribute meaningfully to INSA's mission of advancing Ethiopia's digital sovereignty and Amharic language technologies.

5.2 RECOMMENDATIONS FOR INSA

While INSA already plays a crucial role in advancing Ethiopia's cybersecurity and technology ecosystem, my experience within the **Technology and Innovation Division** gave me insights into areas where improvements or new initiatives could further strengthen its mission. Below are my recommendations:

5.2.1 EXPAND SUPPORT FOR LOW-RESOURCE LANGUAGE TECHNOLOGIES

While INSA already plays a vital role in Ethiopia's cybersecurity and digital transformation efforts, my experience revealed several areas where further initiatives could strengthen its impact:

5.2.2 STRENGTHEN DATA INFRASTRUCTURE

One of the challenges I encountered was the lack of standardized and digitized legal documents. INSA could collaborate with institutions like the Ministry of Justice and Parliament to create a centralized, machine-readable legal database. Such a resource would enhance not only chatbot systems but also legal research, e-governance, and public services.

5.2.3 ESTABLISH AN AI RESEARCH AND INNOVATION LAB

A dedicated research hub within INSA could accelerate Ethiopia's AI ecosystem by focusing on:

- Low-resource NLP models.
- AI-driven cybersecurity solutions.
- Legal technology applications.
- Responsible and ethical AI practices.

This lab could also serve as a collaborative platform for universities, startups, and international partner.

5.2.4 ENHANCE COLLABORATION WITH UNIVERSITIES

Expanding academic-industry partnerships would benefit both INSA and higher education institutions. Initiatives such as joint research projects, mentorship programs, and internships can help nurture the next generation of AI and cybersecurity professionals.

5.2.5 IMPROVE DEPLOYMENT AND ACCESSIBILITY OF AI APPLICATIONS

To maximize social impact, AI prototypes should be deployed as user-friendly platforms accessible to citizens, legal practitioners, and government agencies. For instance, the Amharic Legal Chatbot could be released as a mobile app or integrated into government service portals.

5.2.6 INVEST IN CONTINUOUS TRAINING AND CAPACITY BUILDING

Since technology evolves rapidly, INSA staff and interns would benefit from **regular training programs** in areas such as cloud computing, advanced machine learning, cybersecurity trends, and software deployment. This will ensure that Ethiopia remains competitive in the digital transformation journey.

5.2.7 FOCUS ON ETHICAL AND RESPONSIBLE AI

As Ethiopia advances in AI, it is critical to ensure **responsible AI practices**. INSA could take the lead in drafting **AI ethics guidelines** that safeguard against misuse, bias, or misinformation, especially in sensitive domains such as law, governance, and security

5.3 CLOSING STATEMENT

Overall, my internship at INSA was a powerful learning journey that bridged theory and practice while contributing to a project of national importance. The technical skills, professional values, and experiences I acquired will serve as a strong foundation for my future career in technology and research.

At the same time, the recommendations outlined above—expanding language technologies, strengthening data infrastructure, establishing research labs, deepening academic partnerships, improving deployment, building capacity, and ensuring ethical AI—can further enhance INSA's leadership in Ethiopia's digital transformation.

By continuing to innovate, collaborate, and invest in human capital, INSA has the potential to play a decisive role in shaping Ethiopia's future as a digitally empowered and knowledge-driven society.

REFERENCES

Alemayehu, M., & Gashaw, T. (2022). Natural language processing for Amharic: Current trends and future directions. *Ethiopian Journal of AI Research, 5*(2), 45–62.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). Association for Computational Linguistics.

Federal Democratic Republic of Ethiopia. (1960). *Civil Code Proclamation No. 165/1960*. Addis Ababa: Berhanena Selam Printing Enterprise.

Federal Democratic Republic of Ethiopia. (1995). *The Constitution of the Federal Democratic Republic of Ethiopia*. Federal Negarit Gazeta.

Federal Democratic Republic of Ethiopia. (2004). *Criminal Code Proclamation No. 414/2004*. Federal Negarit Gazeta.

Flask Developers. (2024). *Flask web framework documentation*. https://flask.palletsprojects.com

Google. (2025). *Gemini API documentation*. https://ai.google.dev

Hugging Face. (2025). *Transformers documentation*. https://huggingface.co/docs

Information Network Security Administration (INSA). (2024). *About INSA*. https://www.insa.gov.et

Ministry of Innovation and Technology. (2020). *Digital Ethiopia 2025: A national digital transformation strategy*. Addis Ababa.

Pandas Development Team. (2024). *Pandas documentation*. https://pandas.pydata.org/docs

Qdrant. (2025). *Qdrant vector database documentation*. https://qdrant.tech/documentation

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP* (pp. 3982–3992). Association for Computational Linguistics.

Scikit-learn Developers. (2024). *Scikit-learn user guide*. https://scikit-learn.org

World Bank. (2020). *Ethiopia digital foundations project*. Washington, DC: The World Bank Group.

APPENDIX

Appendix A: Internship Timeline

- **duration**: February 8, 2025 - June 15, 2025
- **Organization**: Information Network Security Administration (INSA)
- **Division**: Technology and Innovation Division – Computing and Data Analytics Department
- **Section**: Data Analytics (AI & Machine Learning)

Appendix B: Tools and Technologies Used

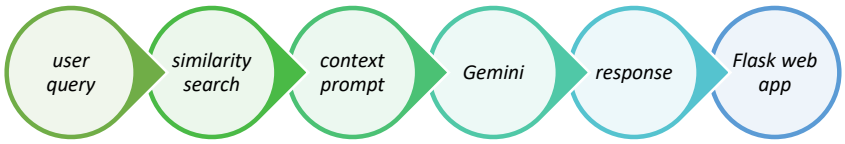
- **Programming Languages**: Python, HTML, CSS
- **Frameworks & Libraries**: Flask, Hugging Face Transformers, Scikit-learn, Pandas, NumPy
- **Machine Learning Models**: paraphrase-multilingual-mpnet-base-v2, transformer-based LLMs (Gemini)
- **Database**: Qdrant / FAISS Vector Database
- **Development Environments**: Google Colab, Jupyter Notebook, VS Code
- **Version Control**: Git & GitHub

Appendix C: Sample Legal Documents Processed

Examples of Ethiopian legal documents used in the project:

1. **Proclamations** – e.g., Civil Code Proclamation, Criminal Code Proclamation.
2. **Regulations** – Government-issued regulations affecting civil and commercial law.
3. **Directives** – Specific operational guidelines from Ethiopian authorities.

Appendix D: Project Architecture Diagram



Appendix E: Sentiment Analysis Sub-Project Summary

- **Goal**: Classify Amharic text into positive, negative, or neutral categories.
- **Techniques**: Data preprocessing, feature extraction, supervised machine learning.
- **Outcome**: Prototype sentiment classifier that achieved promising accuracy, providing insights into text classification in low-resource languages.

Appendix F: Acknowledgment of Training Sessions at INSA

During the internship, I participated in training sessions covering:

- Python for Machine Learning
- Data Cleaning and Preprocessing Techniques

- Vector Databases for NLP
- Flask Web Development Basics
- Ethical Use of AI in Legal and Government Systems