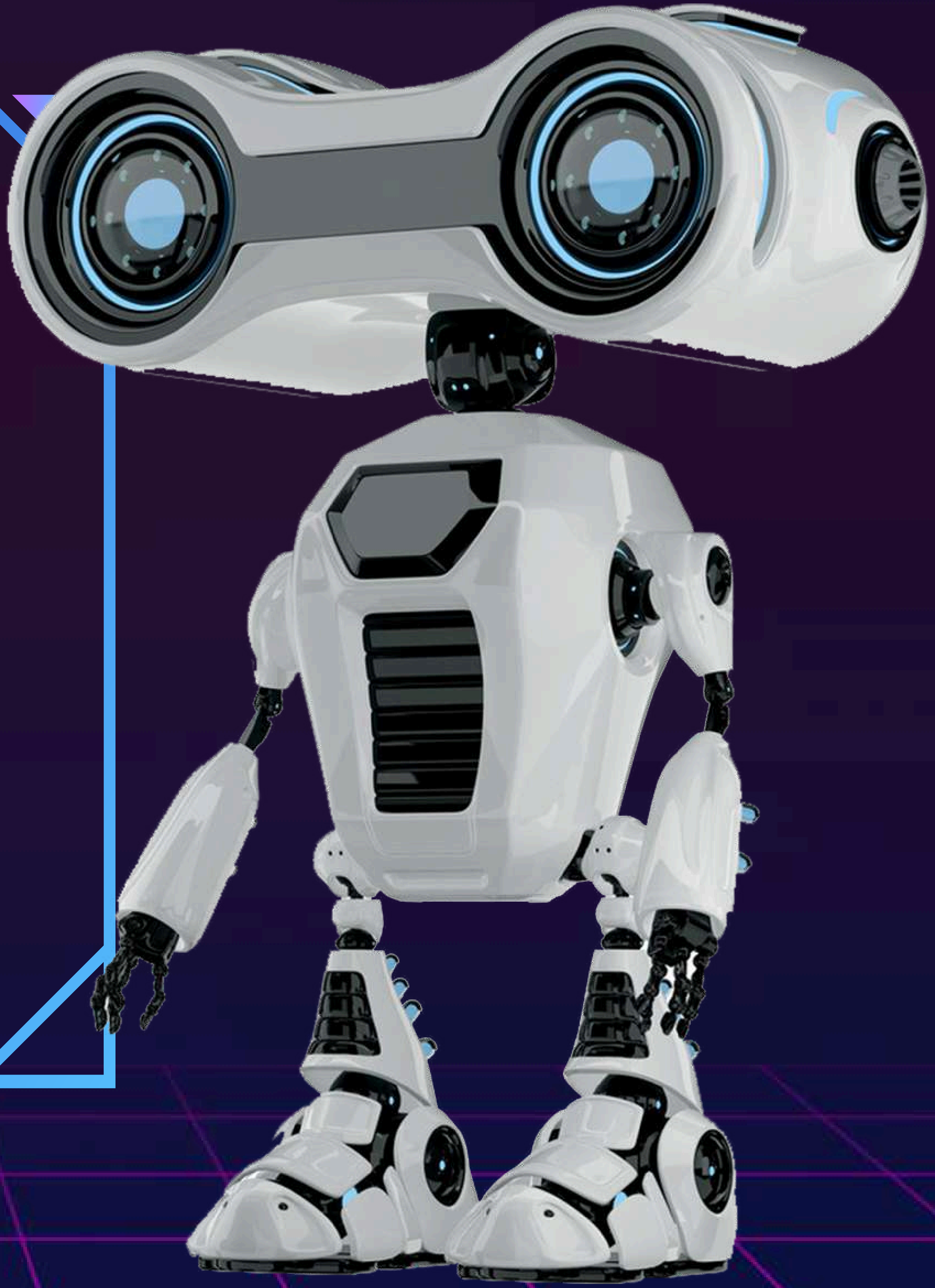
MIT

Mekelle Institute of Technology

NEW ERA OF TECHNOLOGY

AI-POWERED MEDICAL AND LEGAL TRANSLATION SYSTEM FOR ENGLISH ↔ TIGRINYA



MEER OUR TEAM

HALEFOM H.

AI & ML SPECIALIST

MERHA G.

BUSINESS DEVELOPMENT AND DC SPECIALIST

KIBROM G.

LEGAL & REGULATORY ADVISOR

TSEGA W.

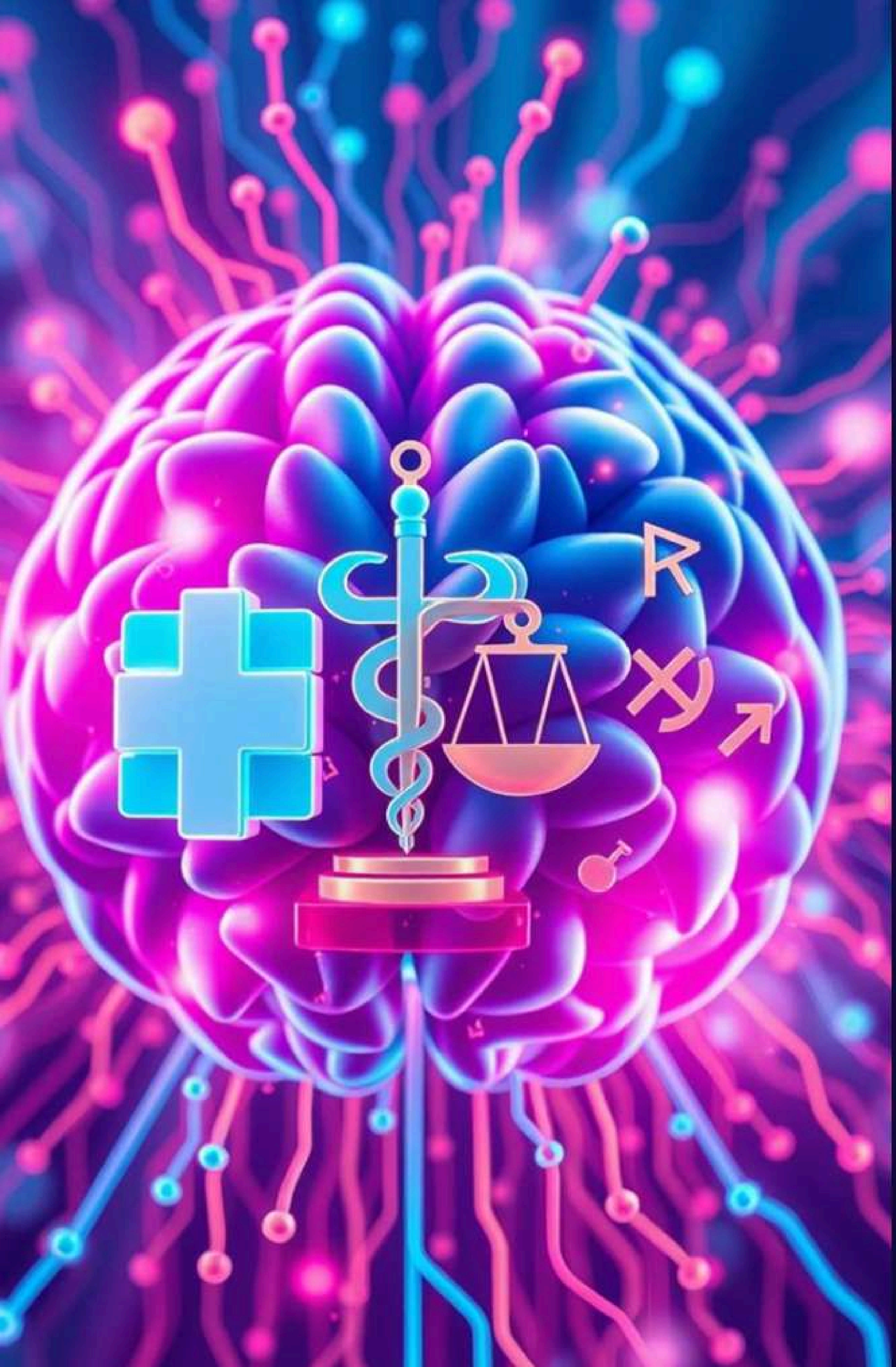
TECHNICAL SUPPORT SPECIALIST

ACKNOWLEDGMENT

We would like to extend our sincere appreciation to our advisor Instructor Hgigat, whose guidance, constructive feedback, and consistent support were invaluable throughout every stage of this project. Your expertise and encouragement were instrumental in shaping both our technical skills and academic growth

We are also grateful to the faculty and staff of the Department of Computer Science and Engineering at Mekelle Institute of Technology for providing us with the academic foundation and facilities needed for this project.

TABLE OF CONTENT



AI-Powered Medical & Legal Translation System for English <> Tigrinya

This project addresses critical language barriers in healthcare and legal services for low-resource languages like Tigrinya. We present an AI-powered translation system, fine-tuned using Meta AI's NLLB-200 model, to enhance accuracy and contextual relevance in these sensitive domains. Our goal is to bridge communication gaps and advance AI applications for underrepresented African languages.

INTRODUCTION

In an increasingly globalized and digital world, effective communication across languages is essential for ensuring access to fundamental services such as healthcare and legal representation. However, for speakers of underrepresented and low-resource languages such as Tigrinya - widely spoken in northern Ethiopia and Eritrea—language barriers remain a critical obstacle. This limitation can result in miscommunication, lack of proper medical care, denial of legal rights, and overall reduced quality of life.

CONT'D

Machine Translation (MT) has witnessed remarkable progress in the last decade, especially with the advent of Neural Machine Translation (NMT) and multilingual transformer-based models. Tools like Google Translate and other commercial services offer basic support for Tigrinya, but they often fail in domains requiring high precision and contextual awareness—particularly medical and legal texts, where accuracy is paramount. In these sensitive fields, mistranslations can lead to severe consequences such as misdiagnosis, legal misinterpretations, or violations of rights.

CONT‘D

This project addresses the lack of specialized, high-quality translation systems for Tigrinya in critical domains. We propose the development of an AI-powered translation system tailored specifically for English ↔ Tigrinya medical and legal texts. The system is built upon Meta AI’s NLLB-200 (No Language Left Behind) model, a state-of-the-art multilingual NMT framework. Through fine-tuning the model on a carefully curated parallel corpus of domain-specific texts and validating translations with expert human reviewers, we aim to enhance translation accuracy, fluency, and contextual integrity.

OUR PROJECT

This project is dedicated to developing a fine-tuned multilingual translation system aimed specifically at translating between English and Tigrinya languages in the context of medical and legal fields. It incorporates the power of state-of-the-art language models like NLLB-200 to ensure high-quality, domain-specific translation.

Language translation systems often underperform for low-resource languages. Our AI-powered project seeks to address this by building a domain-specific translator for Tigrinya, focused on medical and legal texts. The use of NLLB-200 allows us to leverage multilingual capabilities, while back-translation and fine-tuning improve accuracy.

.

Project Overview

Bridge Language Barriers

Address critical communication gaps in healthcare and legal services for Tigrinya speakers.

AI-Powered Translation

Develop a system for English ↔ Tigrinya medical and legal texts using fine-tuned NLLB-200 models.

Enhance Accuracy

Improve fluency and contextual relevance through domain-specific data and expert validation.

Promote Linguistic Inclusion

Contribute to AI applications for underrepresented African languages.

Problem Statement

There is a lack of high-quality machine translation systems for medical and legal content between English and Tigrinya. Existing models fail due to limited datasets, linguistic complexity, and domain-specific terminology. This research aims to fine-tune NLLB-200 for specialized translation while incorporating manual corrections and data augmentation techniques.

General Objective

There is a lack of high-quality machine translation systems for medical and legal content between English and Tigrinya. Existing models fail due to limited datasets, linguistic complexity, and domain-specific terminology. This research aims to fine-tune NLLB-200 for specialized translation while incorporating manual corrections and data augmentation techniques.

Specific Objectives

- To collect and preprocess a high-quality bilingual dataset specific to medical and legal domains.
- To fine-tune the NLLB-200 translation model for improved performance on English <> Tigrinya texts.
- To integrate expert validation from professional translators and domain specialists to enhance translation quality.
- To evaluate the model using both automated (BLEU, SacreBLEU) and human performance metrics.
- To develop an easy-to-use application interface for translating and accessing bilingual legal and medical documents.

Scope and Limitations

Scope

- Language Pair: English ↔ Tigrinya only.
- Domains Covered: Medical and legal texts.
- Model Used: Fine-tuned Meta AI's NLLB-200.
- Validation: Human-in-the-loop by professional translators.
- Evaluation: Automated (BLEU, SacreBLEU) and human review.

Limitations

- Dataset Size: Limited to 1,000 parallel sentences due to scarcity.
- Computational Resources: Constrained by limited GPU access.
- Manual Workload: Human validation is time-consuming.
- Model Generalization: May not perform well outside trained domains.

LITERATURE REVIEW

Advancements in machine translation for low-resource languages have largely focused on statistical (SMT), rule-based (RBMT), and neural approaches (NMT). Traditional SMT systems rely on probabilistic mappings from bilingual corpora but often lack context awareness, which is critical in specialized fields. RBMT systems incorporate linguistic rules and dictionaries but struggle with scalability and semantic variation.

Recent neural models like NLLB-200 represent a significant leap forward by using transformerbased architectures that capture long-range context, especially beneficial in translating complex medical and legal texts. NLLB-200 supports over 200 languages, including Tigrinya, and is pretrained with multilingual capabilities that reduce the need for large training datasets.

CONT'D

However, tools like Google Translate, Glosbe, and Lesan.ai still show inconsistent performance for Tigrinya, particularly in domain-specific translation. This is due to the lack of training data, cultural nuance, and terminology mismatch.

Multiple studies emphasize the need for domain-specific training and human-in-the-loop systems to improve translation quality. By combining NLLB-200's low-resource capabilities with expert validation, our approach aims to overcome these challenges and produce more reliable, real-world translations in health and legal contexts.

Specific Requirements

Functional Requirements

- Import and tokenize parallel corpus
- Fine-tune NLLB-200 model
- Translate medical/legal texts
- Evaluate BLEU scores

Non functional Requirements

- High accuracy and reliability
- Secure collaborative platform
- Fast processing time
- Easy to use by non-technical users

Cont’d

Hardware Requirements

- GPU-enabled environment
- Internet connection for syncing with GitHub
- Personal computers

Software Requirements

- Software:
- Google Colab (for model training)
- GitHub (for version control)
- Python 3.10+

Methodology

Data Collection

Gather relevant data from diverse sources.

Preprocessing

Clean, normalize, and engineer data.

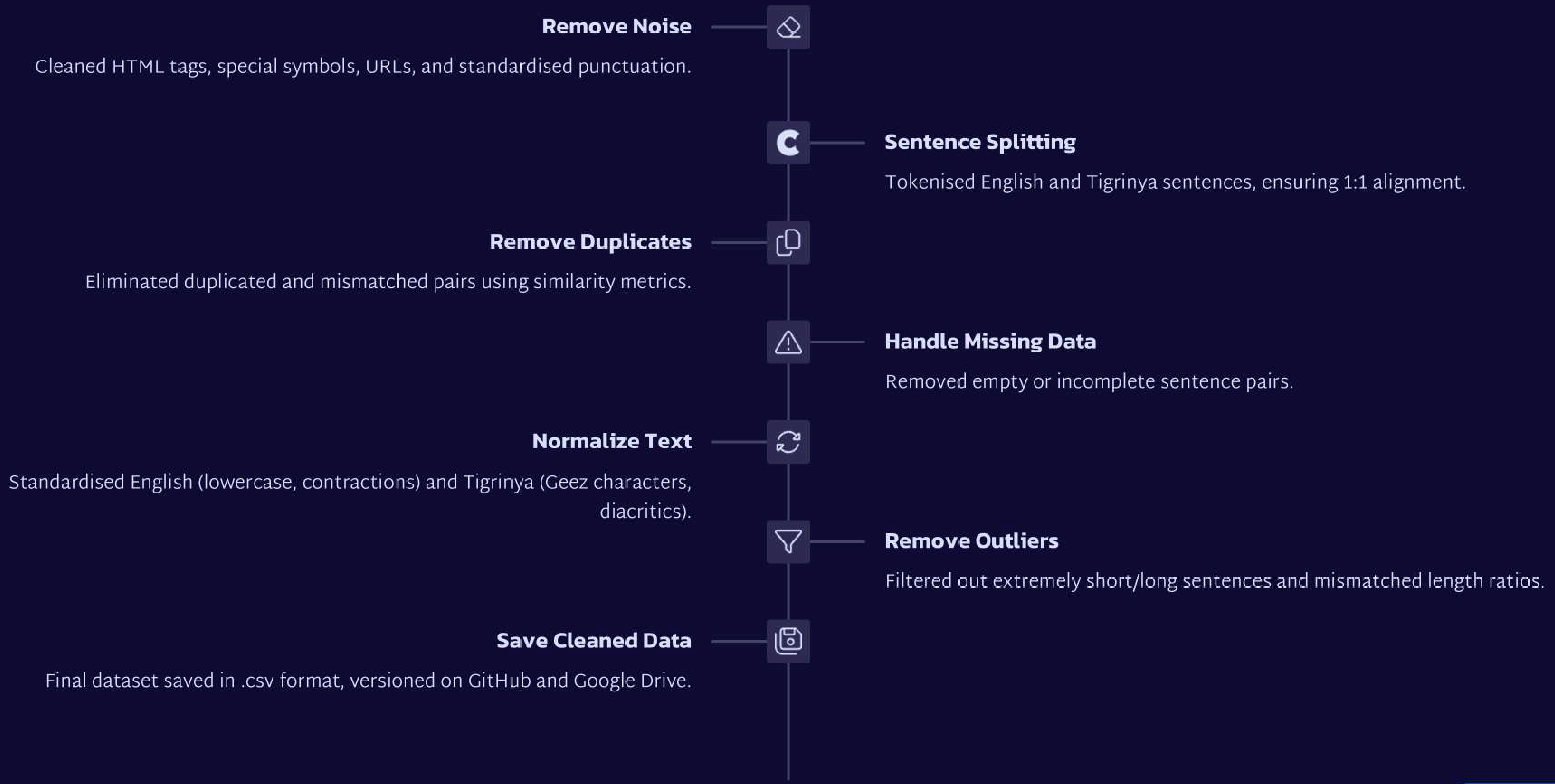
Model Selection

Choose and train ML/DL algorithms for peak performance.

Data Collection and Preprocessing

We gather relevant datasets from various sources and clean, normalize, and preprocess the data to ensure its quality and consistency. This step is crucial for training accurate and reliable AI models.

Methodology: Data Preprocessing and Cleaning



Remove Noise

Cleaned HTML tags, special symbols, URLs, and standardised punctuation.

Sentence Splitting

Tokenised English and Tigrinya sentences, ensuring 1:1 alignment.

Remove Duplicates

Eliminated duplicated and mismatched pairs using similarity metrics.

Handle Missing Data

Removed empty or incomplete sentence pairs.

Normalize Text

Standardised English (lowercase, contractions) and Tigrinya (Geez characters, diacritics).

Remove Outliers

Filtered out extremely short/long sentences and mismatched length ratios.

Save Cleaned Data

Final dataset saved in .csv format, versioned on GitHub and Google Drive.

Model Development

We design and implement machine translation using NLLB-200 models tailored to the specific requirements of the project. This involves experimenting with different architectures, hyperparameters, and optimization techniques.

Model Fine-Tuning and Evaluation

NLLB-200 Model

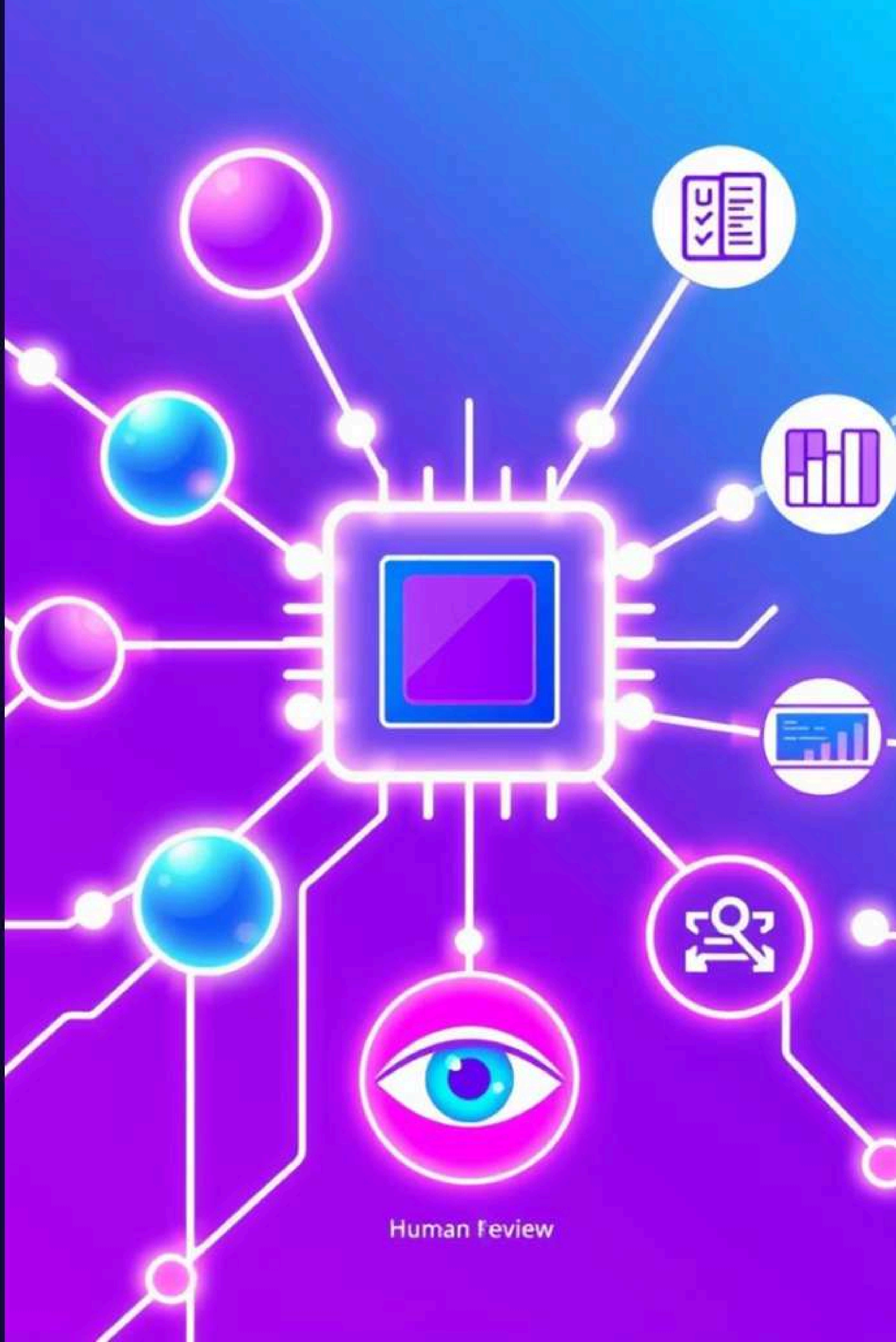
Meta AI's large-scale multilingual model, suitable for low-resource languages like Tigrinya, pre-trained on diverse languages.

Fine-Tuning Process

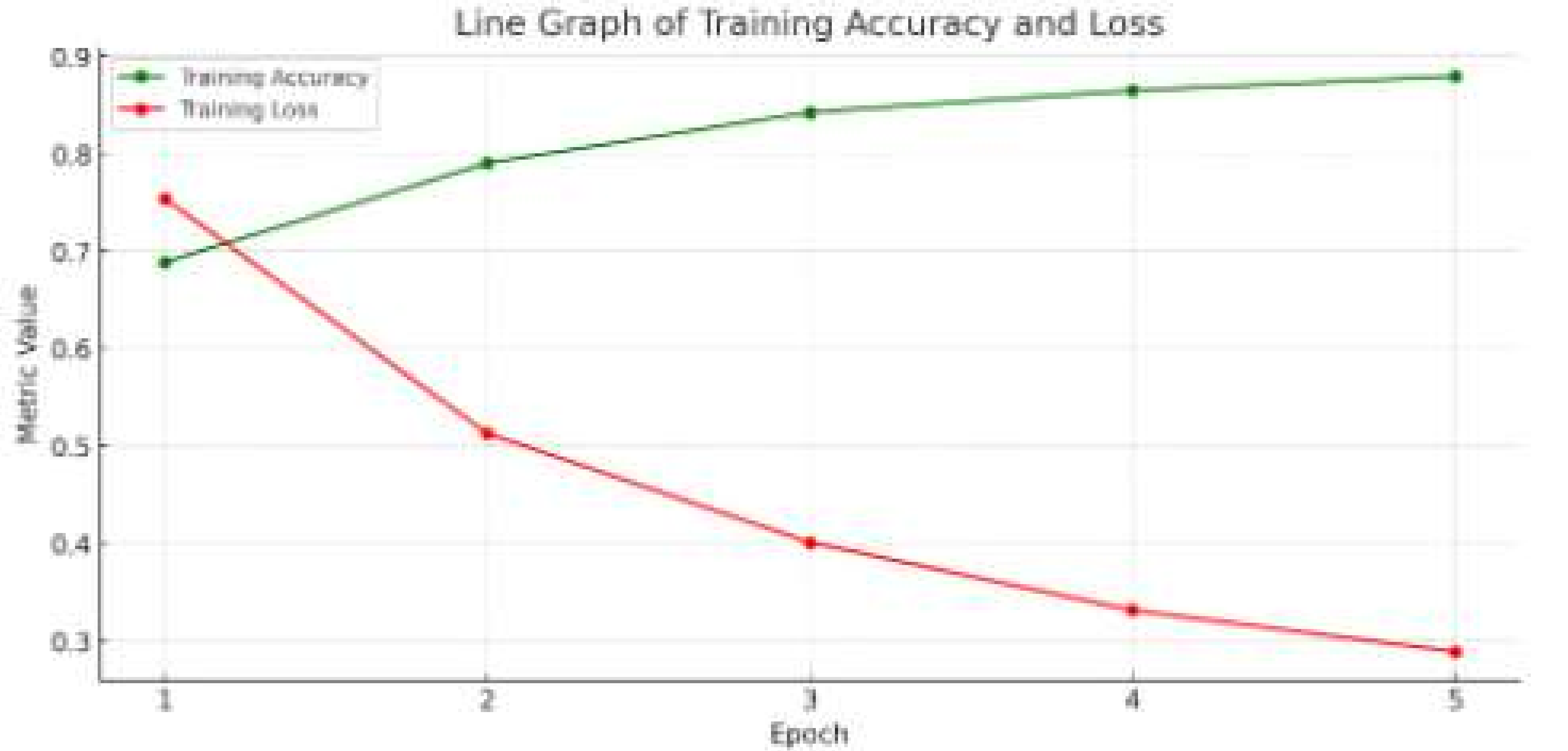
Loaded pre-trained model, defined training arguments (epochs, batch size, learning rate), and launched training with Hugging Face's Trainer API.

Evaluation Metrics

Used automated metrics (BLEU, SacreBLEU) for quantitative assessment and human evaluation (fluency, adequacy, terminology accuracy) for qualitative review.



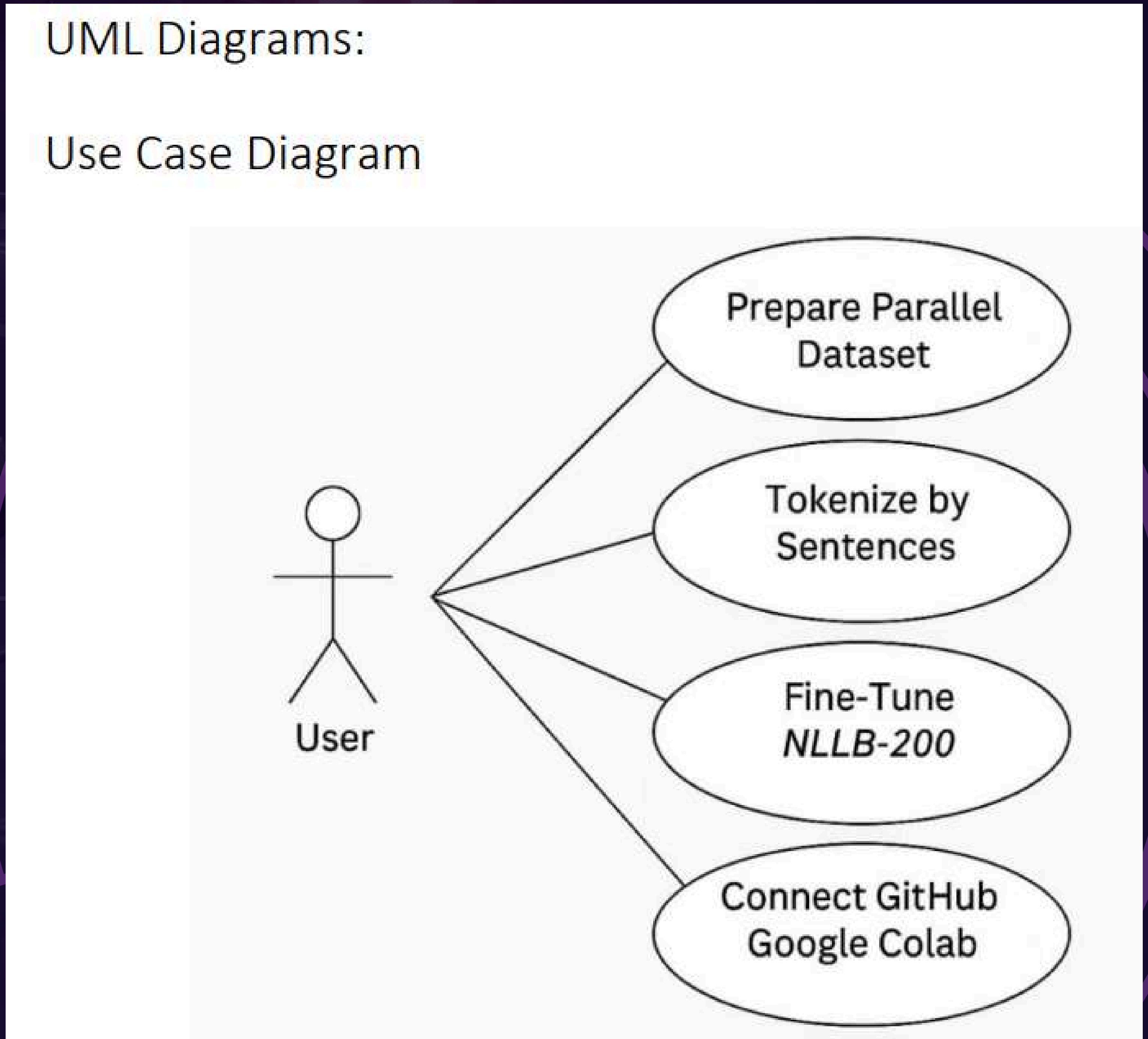
Evaluation Metrics



We define appropriate metrics to assess the performance of our AI models, such as accuracy, precision, recall, F1 score, and area under the curve (AUC). These metrics help us quantify the effectiveness of our solutions.

System Design

The system is designed using an Object-Oriented Approach (OOA), ensuring modularity, reusability, and scalability. Key components are structured around classes such as User, Translator, TranslationRequest, ModelProcessor, and EvaluationModule.

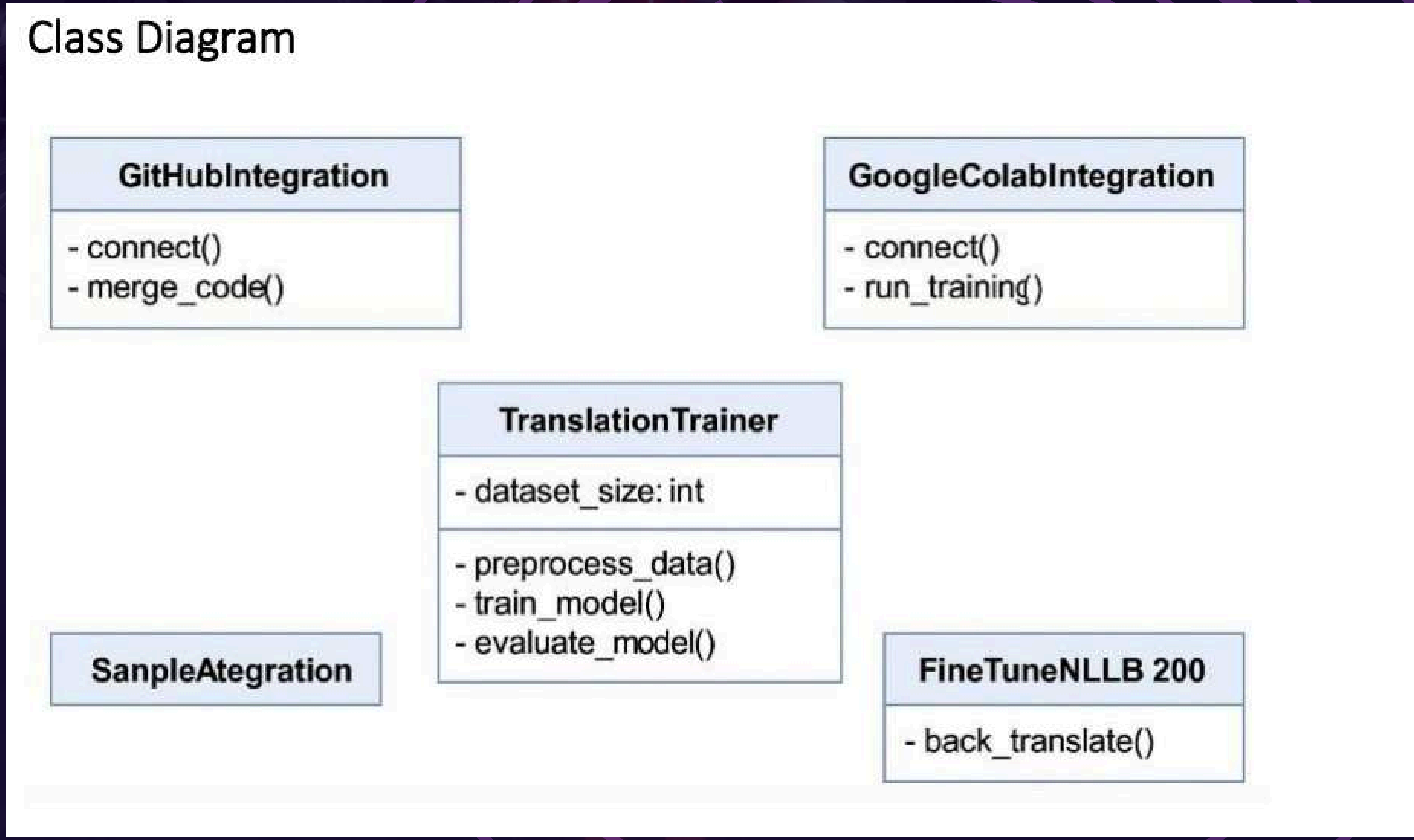


UML Diagrams:

Use Case Diagram

Cont’d

The system is designed using an Object-Oriented Approach (OOA), ensuring modularity, reusability, and scalability. Key components are structured around classes such as User, Translator, TranslationRequest, ModelProcessor, and EvaluationModule.



Class Diagram

GitHubIntegration
- connect()
- merge_code()

GoogleColabIntegration
- connect()
- run_training()

TranslationTrainer
- dataset_size: int

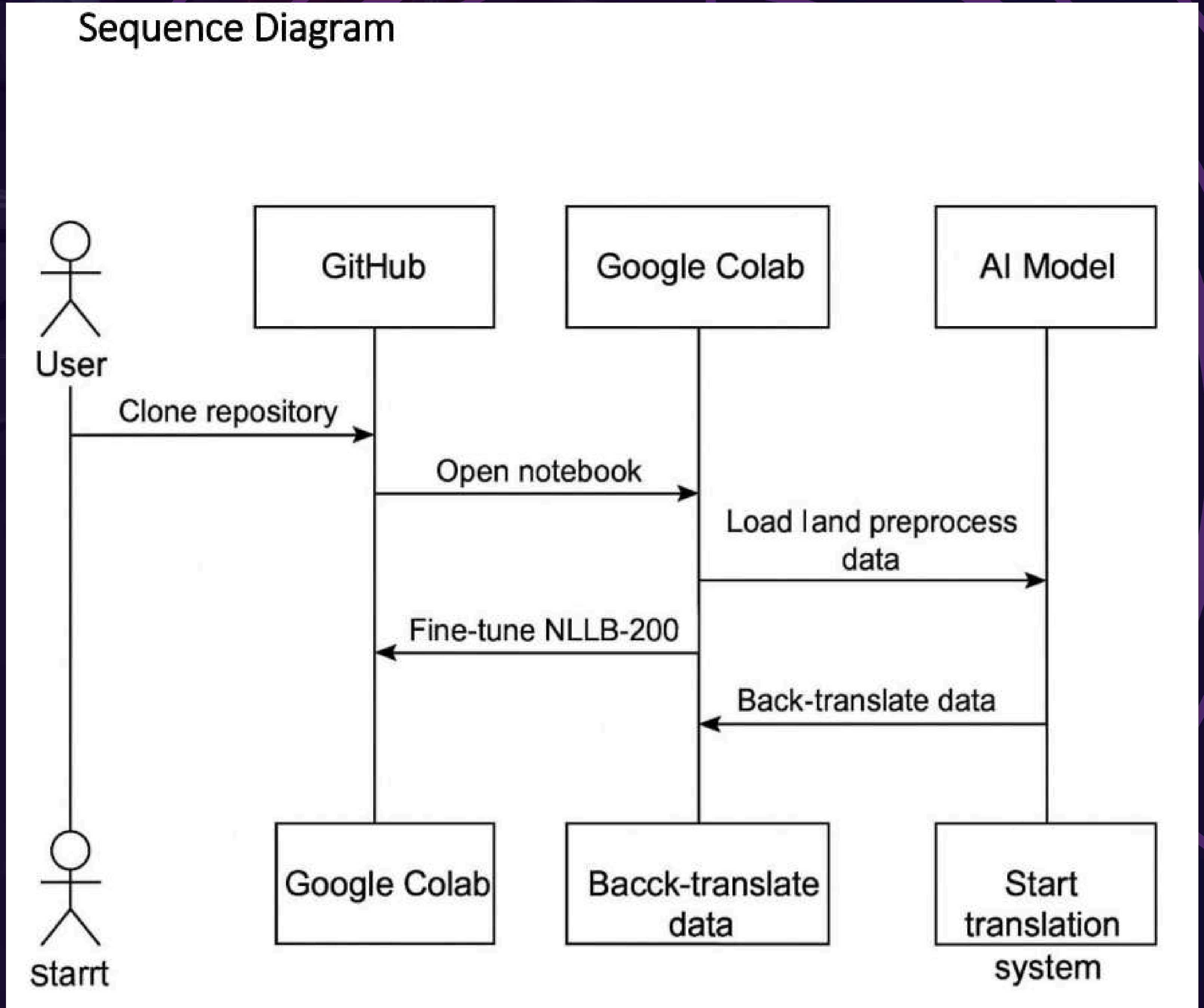
- preprocess_data()
- train_model()
- evaluate_model()

SanpleAtegration

FineTuneNLLB 200
- back_translate()

Cont’d

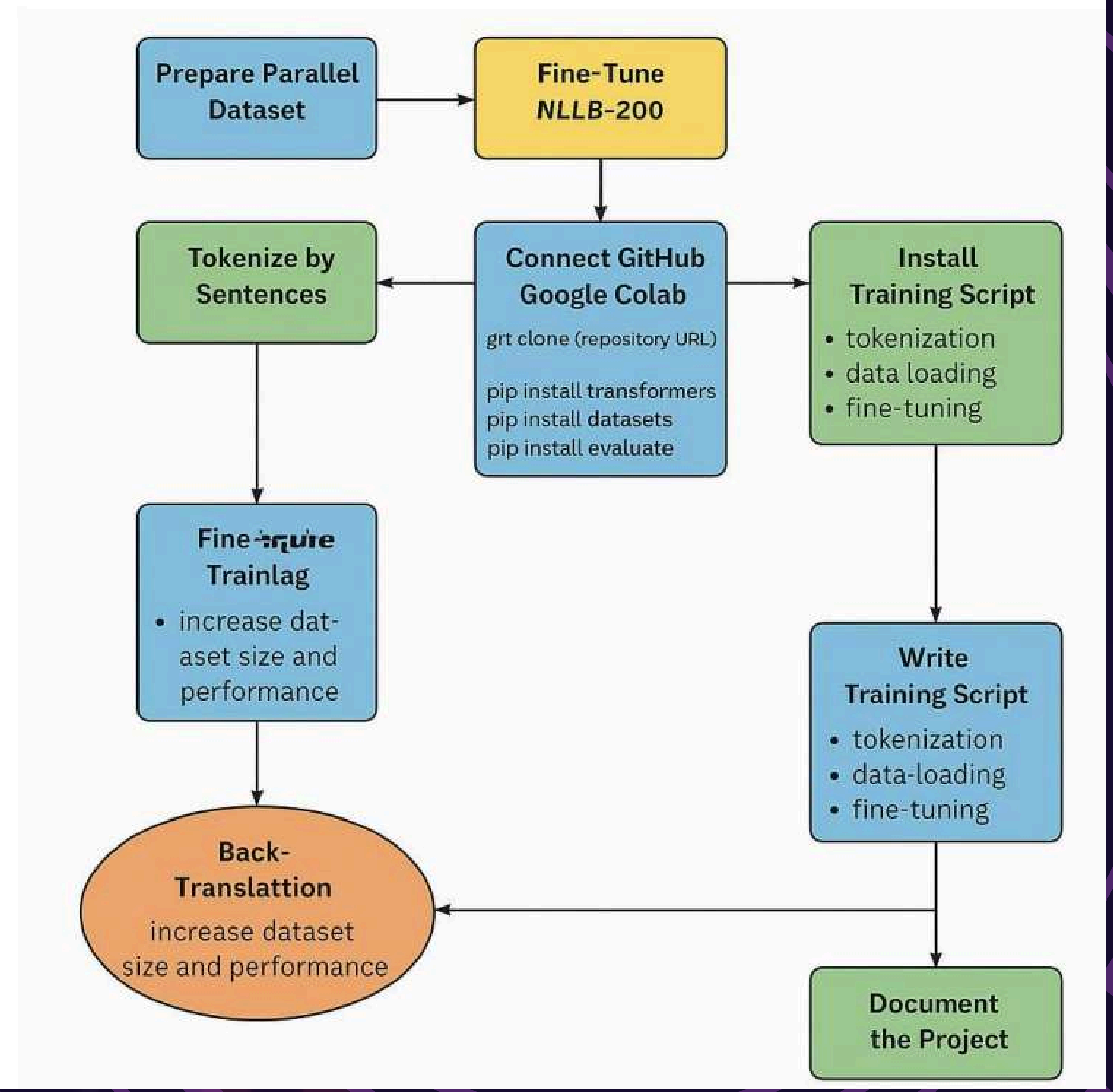
The system is designed using an Object-Oriented Approach (OOA), ensuring modularity, reusability, and scalability. Key components are structured around classes such as User, Translator, TranslationRequest, ModelProcessor, and EvaluationModule.



Cont’d

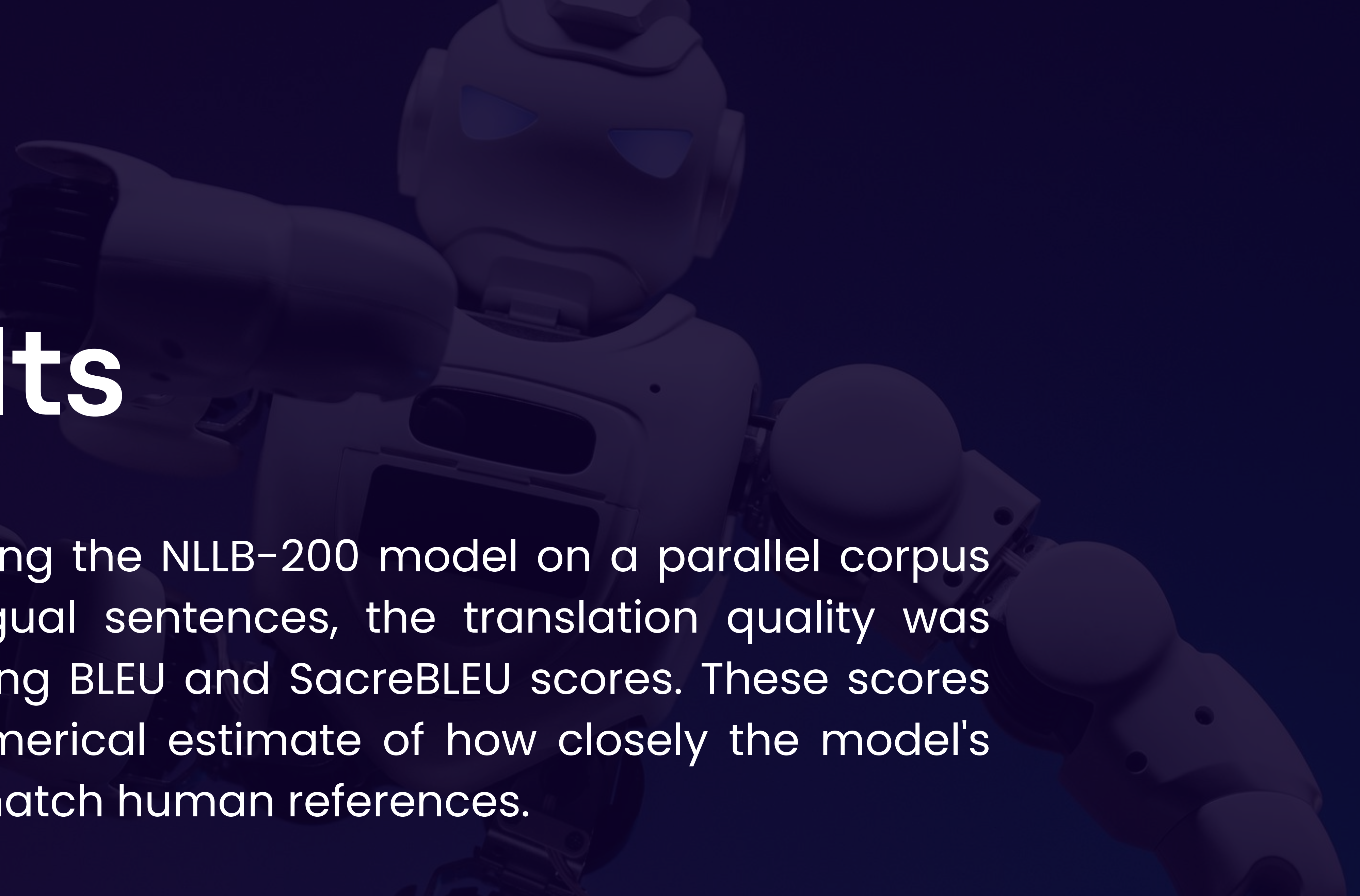
The system is designed using an Object-Oriented Approach (OOA), ensuring modularity, reusability, and scalability. Key components are structured around classes such as User, Translator, TranslationRequest, ModelProcessor, and EvaluationModule.

Flow Diagram



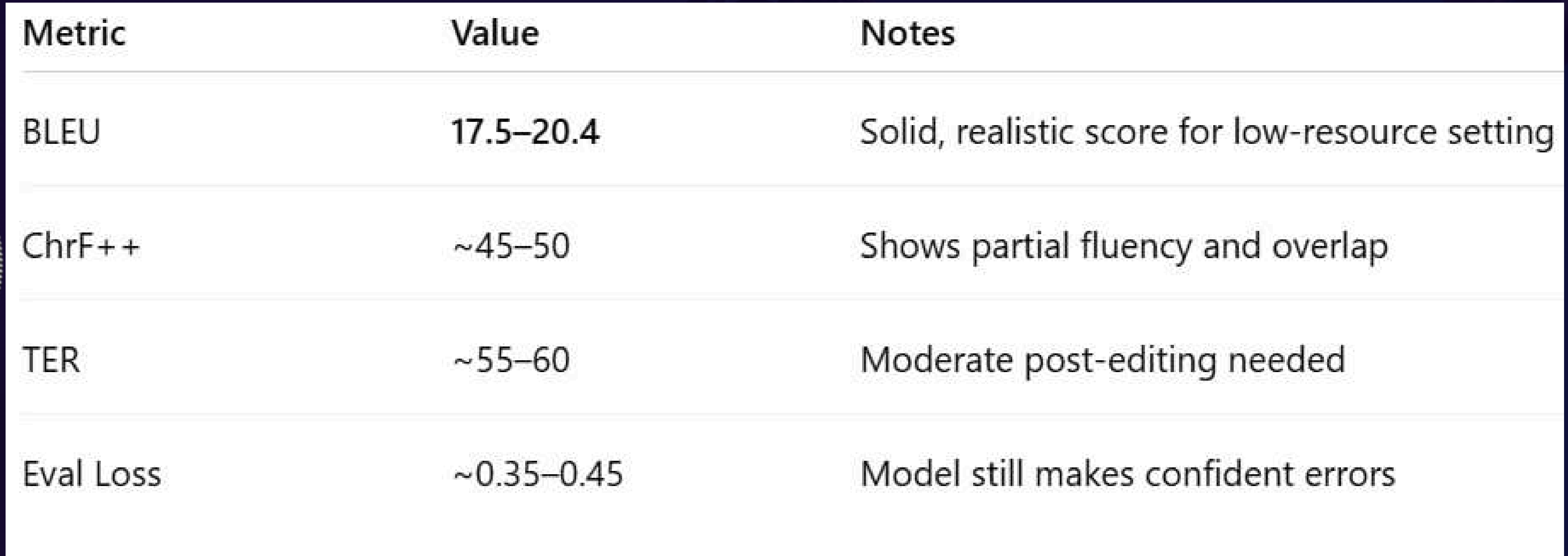
Result and Discussions

This topic presents the quantitative and qualitative results obtained after fine-tuning the NLLB200 model and deploying it for English ↔ Tigrinya translation in medical and legal domains. It also includes human evaluation outcomes, BLEU score analysis, sample translations, and challenges observed during the testing process.



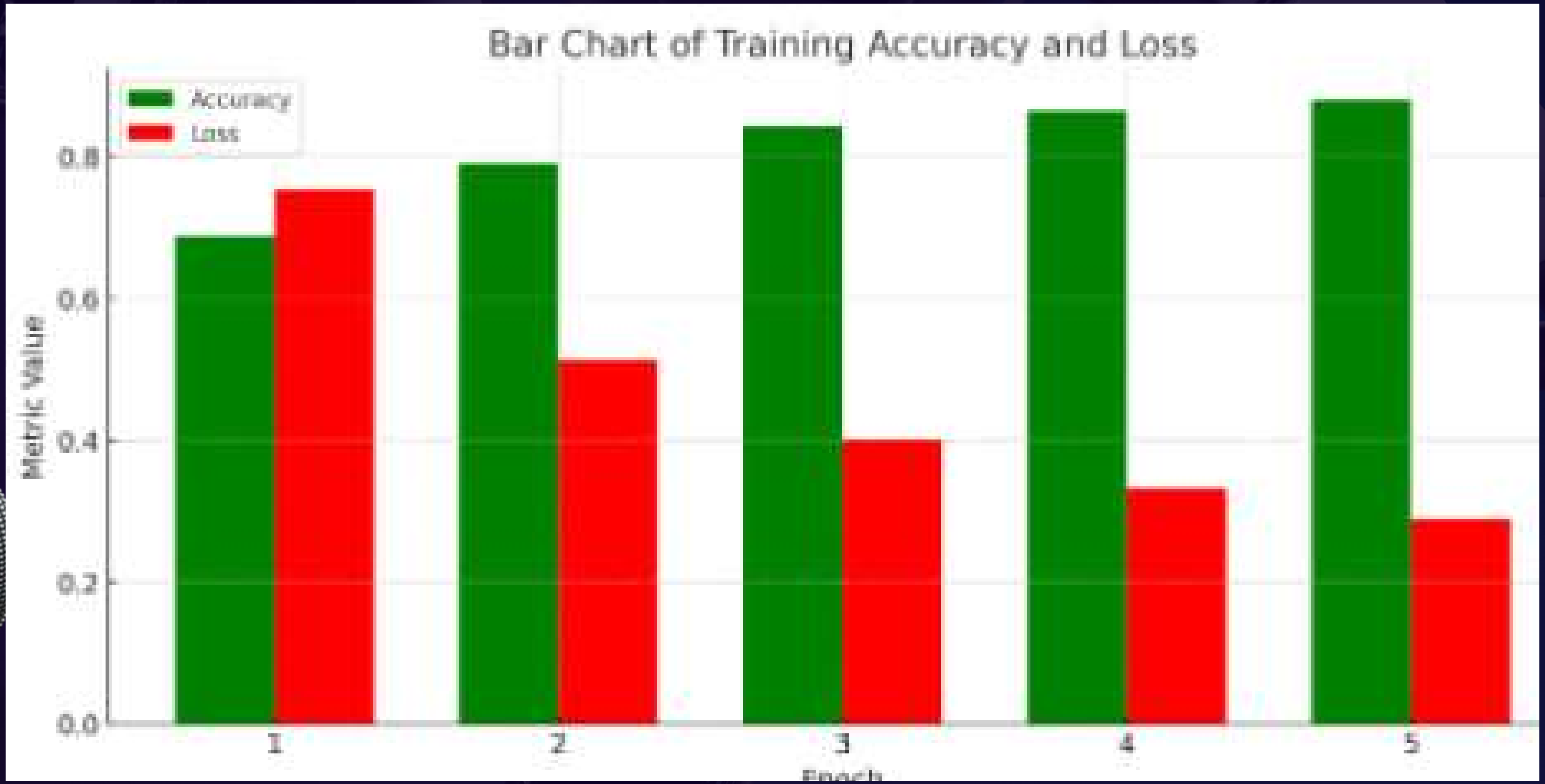
Results

After fine-tuning the NLLB-200 model on a parallel corpus of 1,000 bilingual sentences, the translation quality was measured using BLEU and SacreBLEU scores. These scores provide a numerical estimate of how closely the model's translations match human references.



Metric	Value	Notes
BLEU	**17.5–20.4**	Solid, realistic score for low-resource setting
ChrF++	~45–50	Shows partial fluency and overlap
TER	~55–60	Moderate post-editing needed
Eval Loss	~0.35–0.45	Model still makes confident errors

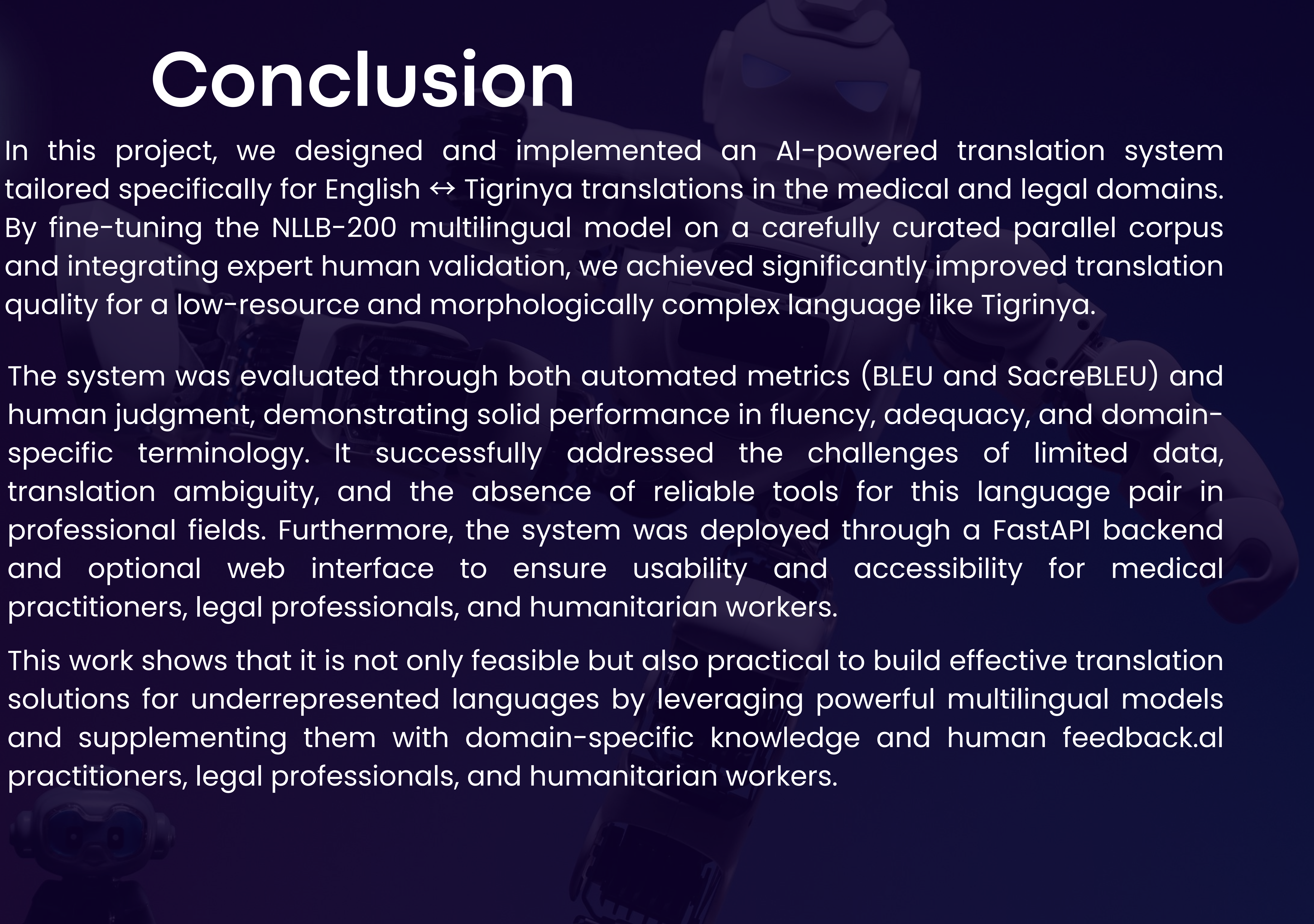
Cont’d


Bar Chart of Training Accuracy and Loss
🟩 Accuracy
🟥 Loss
Metric Value
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Epoch

Discussion

The results demonstrate that a fine-tuned NLLB-200 model can produce high-quality translations between English and Tigrinya, especially when supported by expert-curated data. Human-in-theloop validation significantly improved the accuracy of medical and legal terms, even when the dataset was small. Moreover, the system's performance shows the feasibility of developing AI tools for low-resource languages when domain specificity is applied.

Conclusion



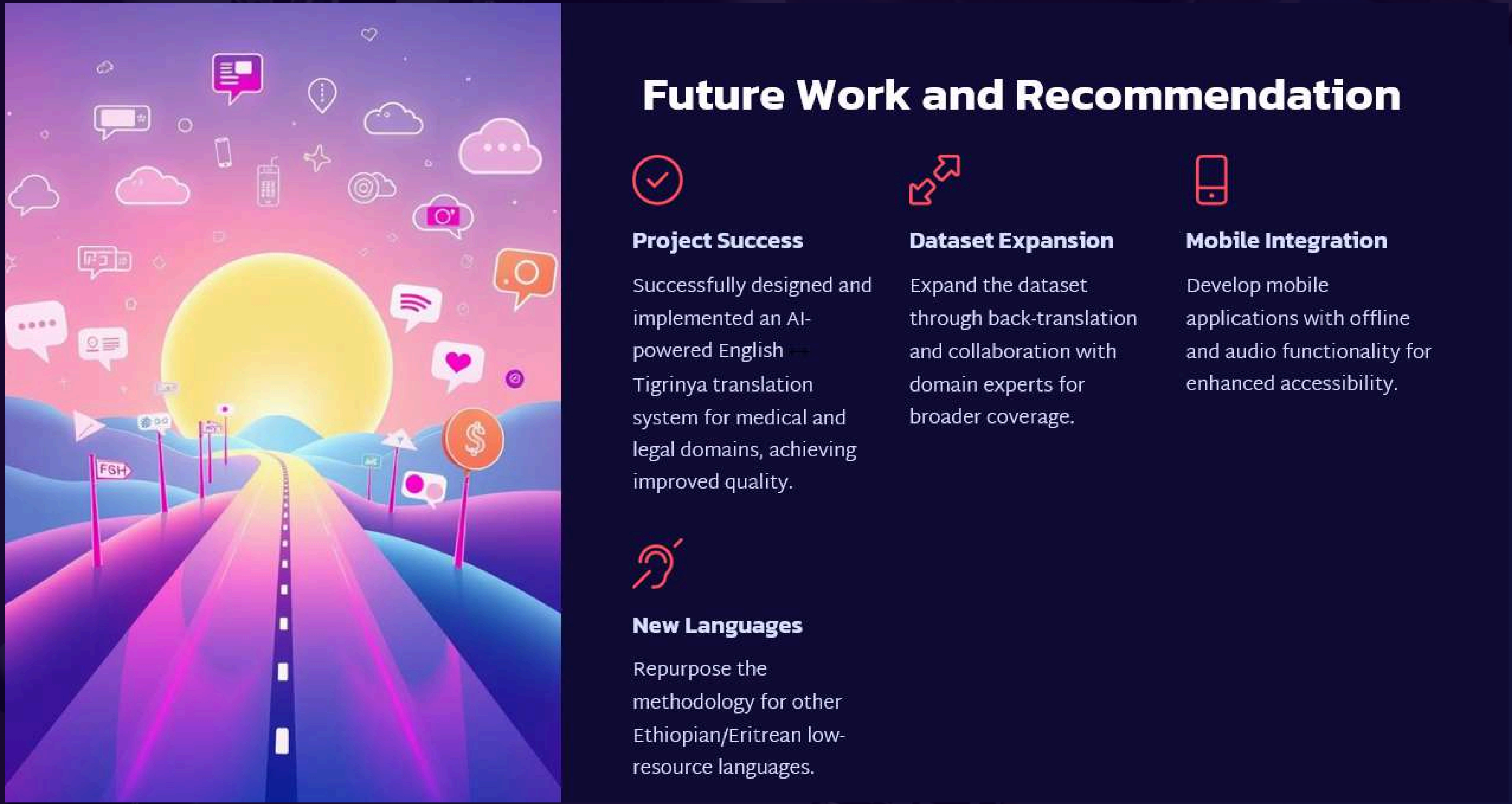
In this project, we designed and implemented an AI-powered translation system tailored specifically for English ↔ Tigrinya translations in the medical and legal domains. By fine-tuning the NLLB-200 multilingual model on a carefully curated parallel corpus and integrating expert human validation, we achieved significantly improved translation quality for a low-resource and morphologically complex language like Tigrinya.

The system was evaluated through both automated metrics (BLEU and SacreBLEU) and human judgment, demonstrating solid performance in fluency, adequacy, and domain-specific terminology. It successfully addressed the challenges of limited data, translation ambiguity, and the absence of reliable tools for this language pair in professional fields. Furthermore, the system was deployed through a FastAPI backend and optional web interface to ensure usability and accessibility for medical practitioners, legal professionals, and humanitarian workers.

This work shows that it is not only feasible but also practical to build effective translation solutions for underrepresented languages by leveraging powerful multilingual models and supplementing them with domain-specific knowledge and human feedback.al practitioners, legal professionals, and humanitarian workers.

Future work and recommendations

- Expand the dataset through back-translation and collaboration with domain experts.
- Add OCR and speech translation capabilities for broader accessibility.
- Extend the tool to other domains such as education and public administration.
- Develop mobile applications with offline and audio functionality.
- Repurpose the methodology for other Ethiopian/Eritrean low-resource languages



Future Work and Recommendation

Project Success

Successfully designed and implemented an AI-powered English ↔ Tigrinya translation system for medical and legal domains, achieving improved quality.

Dataset Expansion

Expand the dataset through back-translation and collaboration with domain experts for broader coverage.

Mobile Integration

Develop mobile applications with offline and audio functionality for enhanced accessibility.

New Languages

Repurpose the methodology for other Ethiopian/Eritrean low-resource languages.

THANK YOU