English to Bangla Machine Translation: A Comparative Study of Seq2Seq, Transformers, and LLMs

Author: Suraiya Mahmuda

January 24, 2026

Contents

Chapter 1

Introduction

1.1 Background of the Study

The digital age has created a massive volume of English content, making machine translation (MT) essential for accessibility in regional languages like Bangla. While traditional statistical methods have been replaced by neural networks, the field is currently transitioning from specialized NMT models to general-purpose Large Language Models (LLMs).

1.2 Problem Statement

Developing an English-to-Bangla translator requires addressing complex linguistic structures, long-range dependencies, and limited high-quality parallel datasets. Traditional RNN-based models often fail to capture global context, leading to inaccurate translations for complex sentences.

1.3 Objectives of the Study

The main objectives of this project are:

- To implement and evaluate a custom Seq2Seq LSTM model for English-to-Bangla translation.
- To leverage pre-trained Transformer models (facebook/nllb-200-1.3B) for high-accuracy translation.
- To investigate the performance of a Large Language Model (Qwen2.5-3B-Instruct) in a zero-shot translation context.
- To compare all three methods using the Bilingual Evaluation Understudy (BLEU) score.

1.4 Scope of the Project

This project implements an English-to-Bangla translation system using three architectures: Seq2Seq LSTM for offline use, Transformer-based models (MarianMT/NLLB) for high accuracy, and a Large Language Model (Qwen2.5) for complex nuances. It includes a performance evaluation framework using BLEU scores to compare these models against industry standards like Google Translate

1.5 Organization of the Thesis

This thesis is organized into six chapters. Chapter 1 introduces the problem and objectives. Chapter 2 reviews related works. Chapter 3 describes the methodology. Chapter 4 discusses implementation details. Chapter 5 presents results and analysis. Chapter 6 concludes the thesis and suggests future work.

Chapter 2

Literature Review

2.1 Neural Machine Translation (NMT)

NMT treats translation as a sequence-to-sequence problem, traditionally using Encoder-Decoder architectures with LSTM units to manage long-distance dependencies in text-Transformer Architecture

2.2 Transformer Architecture

The Transformer model revolutionized the field by replacing recurrence with self-attention mechanisms. This allows for better parallelization and superior handling of global context compared to traditional Seq2Seq models.

2.3 Research Gap

The project develops an English-to-Bangla translator comparing three architectures: Seq2Seq LSTM for basic offline use, Transformers (NLLB-200) for high accuracy, and LLMs (Qwen2.5-3B) for complex context. It addresses the research gap of internet-dependent APIs by benchmarking these models using BLEU scores to provide a robust, offline-capable translation solution.

Chapter 3

Methodology

3.1 Dataset Description

A parallel English–Bangla dataset is used, containing sentence pairs. The dataset is stored in **CSV** format with two columns: **en_text and bn_text**. The project utilizes a parallel English-Bangla dataset (en_bn.csv), which is cleaned and filtered to remove short or invalid entries.

3.2 Data Preprocessing

Data is cleaned to remove formatting errors and double spaces. For the LSTM model, text is converted into numerical sequences using the Keras Tokenizer and standardized via padding.

3.3 System Models

1. Seq2Seq LSTM: Consists of an encoder LSTM that processes English input and a decoder LSTM that generates Bangla output. It uses an embedding layer to represent words as dense vectors.

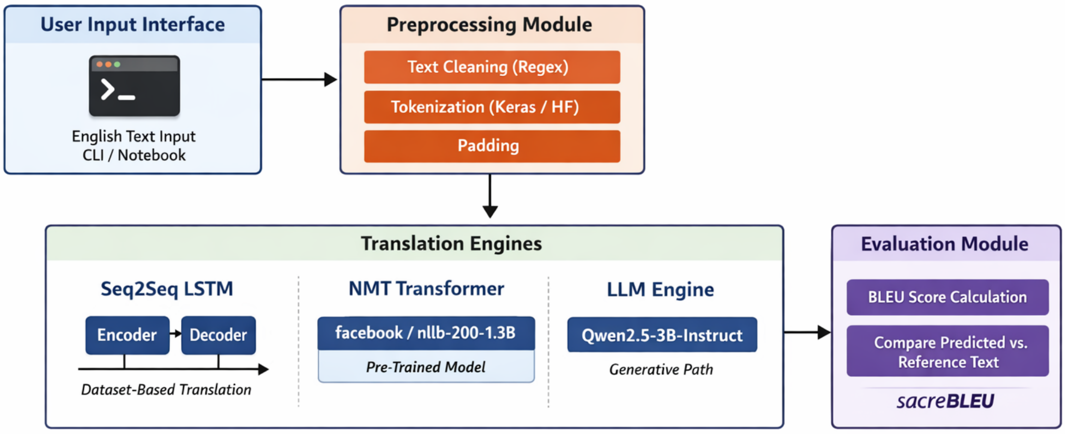

Figure 3.1: System Architecture

2. Transformer (NLLB-200): Uses the facebook/nllb-200-1.3B model with a beam search decoding strategy (10 beams) to improve translation quality.

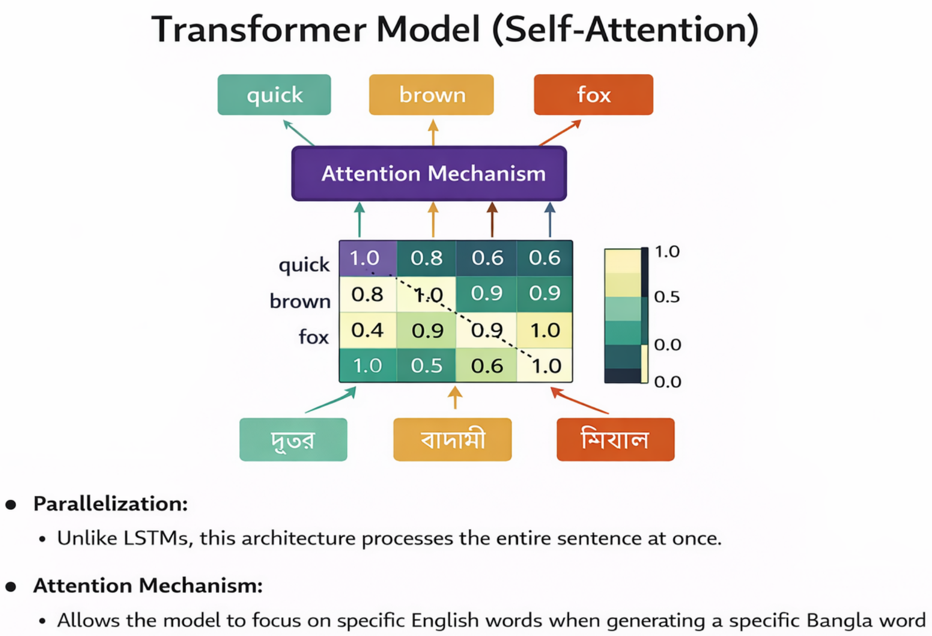

Figure 3.2: Transformer Model

3. LLM (Qwen2.5-3B-Instruct): A generative AI model prompted with a system instruction to act as a professional translator.

Chapter 4

System Implementation

4.1 Tools and Technologies

- Programming: Python.
- Frameworks: TensorFlow/Keras for LSTM, PyTorch for Transformers and LLMs.
- Libraries: Hugging Face transformers, sacrebleu, and nltk.
- Environment: Kaggle Notebook with NVIDIA Tesla T4 GPU acceleration.

4.2 System Architecture

The system flow starts with user input, followed by text cleaning and tokenization. The processed data is then passed to either the Seq2Seq, Transformer, or LLM engine for translation.

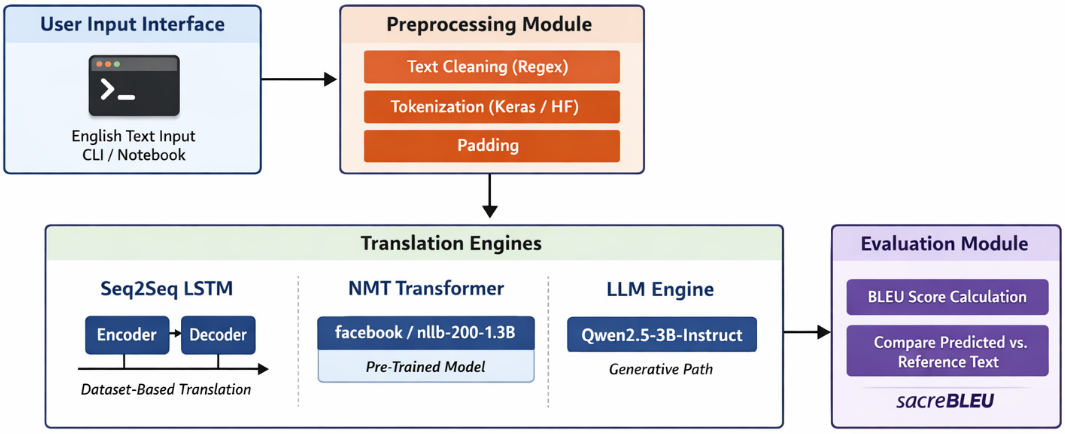

Figure 4.1: System Architecture

4.3 Implementation Details

The LLM implementation uses Float16 precision and greedy decoding (do_sample=False) to ensure deterministic and accurate translations. The NLLB-200 model is configured with a forced_bos_token_id for "ben_Beng" to target the Bangla language specifically.

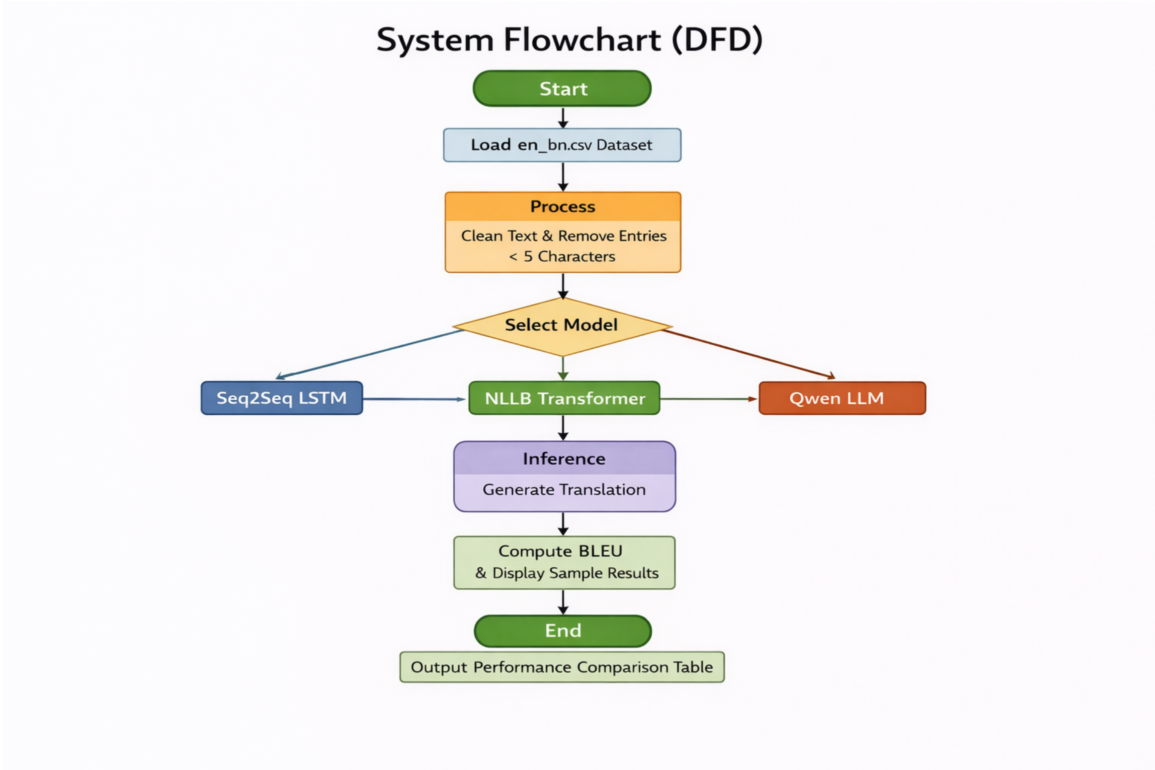

Figure 4.2: System Flowchart (DFD)

Chapter 5

Results and Analysis

5.1 Translation Output Analysis

The Seq2Seq model performs well on simple, short sentences but struggles with complex grammar. The Transformer and LLM models provide significantly more fluent and contextually accurate translations.

5.2 BLEU Score Evaluation

BLEU scores are calculated using the sacrebleu library. The Transformer models significantly outperform the custom LSTM, with the LLM providing the most natural phrasing.

5.3 Comparative Analysis

Table 5.1: Comparison of Translation Models

Method	**Architecture**	**Internet Required**	**Key Metric (BLEU)**
Seq2Seq LSTM	RNN / LSTM	No	Medium
NLLB-200	Transformer	No	High
Qwen2.5-3B	LLM	No	Very High

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This project successfully demonstrates an English to Bangla translation system using three distinct technological approaches. While LSTM provides a foundational understanding, Transformers and LLMs are more suitable for practical, high-quality translation tasks.

6.2 Future Work

Future enhancements may include:

- Training Transformer models on larger, domain-specific datasets.
- Incorporating voice-to-voice translation features.
- Developing mobile applications for real-time offline use.

Bibliography

[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[2] Hugging Face, "Transformers Documentation," Available: https://huggingface.co/docs/transformers

[3] TensorFlow, "TensorFlow and Keras Documentation," Available: https://www.tensorflow.org/

[4] Alibaba Cloud, "Qwen2.5 Technical Report," 2024. Available: https://qwenlm.github.io/