Bengali Natural Language Processing: A Fine-Tuned Approach to Question Answering

By
Debashish Roy
213-15-14771

Md. Rafit Mostofa Noor
213-15-14784

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements forthe **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by

Amir Sohel
Assistant Professor
Department of Computer Science and EngineeringDaffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 16, 2025

APPROVAL

This Project titled **"Bengali Natural Language Processing: A Fine-Tuned Approach To Question Answering"**, submitted by **Debashish Roy (213-15-14771)** and **Md. Rafit Mostofa Noor (213-15-14784)** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **16 September, 2025**.

BOARD OF EXAMINERS

Dr. Bimal Chandra Das (BCD) — **Chairman**
Professor and Dean (in-charge)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Most. Hasna Hena (HH) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Mayen Uddin Mojumdar (MUM) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Nazibur Rahman — **External Examiner**
Technical Lead – Database Administrator
Telenor – Grameen Phone Account

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sazzadur Ahamed, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Mr. Amir Sohel
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Debashish Roy

Debashish Roy
Student ID: 213-15-14771
Department of Computer Science and Engineering
Daffodil International University

Rafit Mostofa Noor

Md. Rafit Mostofa Noor
Student ID: 213-15-14784
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Md. Sazzadur Ahamed**, **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Natural Language Processing** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Bengali is a native language spoken by more than 250 million people all over the world. Building a system that can understand human language and provide meaningful responses is really important in Natural Language Processing (NLP), which is also advancing rapidly. Building a system in Bengali is difficult, as there are scarcity of data, complex morphological and versatile syntax. The project titled "**Bengali Natural Language Processing: A Fine-Tuned Approach to Question Answering**", creates a QA model intended for Bengali Language. We have created a structured general knowledge based Bengali question-answer dataset which is used to fine-tune a pre-trained language model. The proposed system is a transformer based architecture to enhance the comprehension and contextual reasoning capability of Bengali QA system. With a domain specific dataset fine tuning the proposed model can improve performance in understanding of Bengali language. This work also highlights the challenges of handling limited resources and present an approach to generalize other low resource languages. The main goal is to enrich the model's ability to understand and answer questions in Bengali with more accurate and also contributing the advancement of NLP tools for low resource language.

Table of Contents

List of Figures

Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the overview of the work on question answering in Bengali which is low-resource language processing. It describes the goal of the project and employing a multilingual model and the techniques used for productive question-answering.

1.1 Introduction

Natural Language Processing (NLP) is advancing rapidly for building systems which can understand human language and can make meaningful responses. A great number of progress has been made in languages like English and Chinese. But native language like Bengali remain behind in this race of NLP. Specially in question answering (QA research [1]. These lacking of Bengali language highlights the need of efforts to develop resource and system which can address the linguistic complexities.

More than 250 million of people speaks in Bengali all over the world. Thus it is one of the mostly spoken language of the world. In spite of having huge number of speakers, computational resources and NLP applications are still limited and low comparison the other high resource languages. Question answering system specific goal is to generate precise responses to user queries in Bengali language. The development of this system faces various problems because for the lack of qualityful datasets and tools.

One of the difficult tasks of building QA systems in Bengali stems fro more linguistic and technical challenges. These difficulties are datasets, complex morphological structures of the language and its versatile syntactic patterns [2]. These features make complications in preprocessing, dataset design and model training which are basic steps in building effective QA systems.

For these challenges, the project titled “Bengali Natural Language Processing: A Fine-Tuned Approach To Question Answering” aimed to create a structured qualityful dataset and the development of a QA system. This approach ensures the model contribution to the advancement of NLP.

1.2 Motivation

The demand for Bengali Content and highlights the need for specific-language models.Most

of the QA models are good in English but are weak in other languages because of the lack of training datasets.Bengali is one of the largest native language and there is not plenty of resources and researches in NLP .This project is was motivated to reduce the linguistic gap by developing a QA model for Bengali and ensure a great AI representation for low-resources language like Bengali.

Bengali is the seventh most spoken native language around the world. More then 250 million of people speaks Bengali but it is a low-resource category language in terms of NLP research in comparison to English and other regional languages like Hindi or Chinese. The shortage of proper dataset , pre-trained models, and dedicated research efforts has led to the situation which is Bengali speakers do not have enough or equal benefits of modern AI technology. For these reason Bengali is not a developed language of intelligent systems which can process, understand, and generate meaningful responses.

Because of linguistic and technological gap this project was inspired. The basic objective of this project is to build a Qa model which is tailored for the Bengali language, so that we can contribute to the field of low-resource NLP. By completing this project will lays to the basic groundwork for future by providing the immediate need for a functional Bengali QA system. Except the technical contribution, this project also carries a great strong social and cultural motivation. Languageis the primary aspect of identity, and AI representation ensures the Bengali language inclusivity by enabling millions of people to interact with intelligent systems in their mother language.This will make sure that the development in AI is not limited to with some language.

The improvement and development of a Bengali QA model is not only a technical achievement but also will glorify and symbolise the effort to remove linguistic disparity, will also promote digital inclusivity, and will strengthen the presence of Bengali language which is evolving greatly in the world of Artificial Intelligence.

1.3 Objectives

The Primary objectives of this project:

i) To develop a new structure and qualityful Bengali QA dataset
ii) Develop an efficient question-answering system for Bengali language by fine-tuning a pre-trained model.
iii) Compare effectiveness of LLaMA with the existing Bengali NLP models.
iv) Improve accessibility of advanced NLP tools for Bengali users in education, research and real world applications.

v) Contribute to the research community by releasing the dataset and the fine-tune model for future developments.

1.4 Methodology

The research design of this project follows a systematic approach to fulfill the gap in Bengali NLP. Specially in QA system. The procedure starts with identifying the lack of resources, which is followed by a detailed review of existing QA systems and datasets. Then we created a dedicated and well furbished Bengali QA dataset and preprocessed them to ensure the quality and consistency. Then a suitable model for fine-tuning like s BERT, Gemma, LLaMA, GPT-2, and BanglaBERT are selected. The models are trained and evaluated using different standard metrics such as BLEU and Exact Match (EM) scores. The results of this is then compared with the previous work to highlights the improvements and limitations.

Research Design Steps are:

i. Problem identification: Gap in Bengali NLP resources, specially in QA datasets.

ii. Literature Review: Studying existing QA systems and datasets.

iii. Dataset Design: Collection of Bengali questions and answers in JSONL format.

iv. Preprocessing: Remove duplicates, cleaning, tokenization.

v. Model Selection: Fine-Tuning models like BERT, distilBERT, LLaMA, GPT-2, BanglaBERT etc.

vi. Training and Evaluation: Use our dataset to train and evaluate using BLEU, EM score.

vii. Analysis and result Interpretations: Compare with existing works

viii. Conclusion and Future work: Limitations, scalability and improvement suggestions.

Data Collection & Preparation

The dataset for this project is created manually. A total of 15000+ questions-answer pairs are collected. The topic of this are mainly related to geography, history, education and everyday facts related to Bengali speakers. All the entries in the dataset follows the JSON format like {“input” : “question”, “output” : “answer”} / {“input” : "প্রশ্ন", “output" : “উত্তর”}

Table 1.1: Dataset overview

Input (প্রশ্ন)	Output (উত্তর)
কোন আইনের অধীনে রাজশাহী সিটি কর্পোরেশন গঠিত হয়েছে?	রাজশাহী সিটি কর্পোরেশন আইন, ১৯৮৭
কমিশনের সচিব কীভাবে নিযুক্ত হন?	সরকার কর্তৃক নিযুক্ত হন।
নির্বাচন পরিচালনার জন্য সরকার কী প্রণয়ন করে?	বিধি
সিটি কর্পোরেশনের মেয়াদ কত বছর?	পাঁচ বছর

The pairs of the dataset are written using different Bengali structure and a common phrasing styles.This would be helpful for the model to understand the real-world user enquiries. This dataset reflects the challenges in Bengali NLP, including various kind of spelling, sentence length, grammar patterns. We also go through manual verification to ensure that each pair is accurateThis custom dataset provides a solid foundation for fine-tuning transformer-based Bengali models, making the QA system more reliable and effective in practical use.

1.5 Analysis Techniques

To evaluate the developed Bengali system several analysis techniques are used -

i) Dataset Analysis: Statistical analysis is performed to examine the question lengths, answer types, and overall dataset balance.

ii) Model Evaluation: The model will be evaluated using standard metrics like Exact Match (EM) and F1 score. These metrics provide models ability to return accurate and relevant answers.

iii) Error Analysis: Model predictions will be manually reviewed to identify common errors and improvement areas.

1.6 Project Outcome

The prime outcome of this project is to create a high quality Bengali Question-Answering (QA) pair dataset and fine-tuning a model. The dataset is created carefully in jsonl format. The reason behind this format if for advance research in Natural Language Processing (NLP) for low-resource language like Bengali. This process includes a great number of QA pair in different domain like general knowledge, history, literature that ensures a broad application. To continue the linguistic accuracy, each and every entity has been verified for grammatical correctness and contextual relevance.

The most valid and important reason behind to create this dataset is to fine-tune the large language model LLaMA. This model architecture offers a significant importance for developing Bengali QA system. Technically the dataset's jsonl structure ensures seamless integration into deep learning workflows and supports scalable experimentation. Later on it will be released as an open source under a valid public license and will be available for researchers and developers on different platforms like HuggingFace, Kaggle and GitHub.

1.7 Organization of the Report

The thesis is organized into six chapters, each addressing different stages of the project development :

Chapter 1 Introduction: Introduces the background, motivation, objectives, methodology, analysis techniques, project outcome.

Chapter 2 Background: Provides an overview of existing research in the field of Question-Answering, specially in low resource languages like Bengali.

Chapter 3 Research Methodology: Describe the steps in designing and building the dataset including data collection, annotations, formatting (jsonl), preprocessing and etc.

Chapter 4 Implementation and Results: It includes dataset statistics, evaluation metrics (EM, F1), model training results and error analysis.

Chapter 5 Engineering Standard and Design: ChallengesDescribes the engineering problems and solutions has been solved in this project.

Chapter 6 Conclusion: Summarizes the findings and highlights the contributions of the project, discusses about limitations and future directions.

Chapter 2

Background

This chapter gives a comprehensive background in Bengali question answering, which highlights the major challenges of developing NLP tools for low-resource languages like Bengali. QA system have significant upgradation and progress in high resource languages but in terms of Bengali the research is limited for the scarcity of dataset over reliance on multi-lingual model with reduced accuracy and the absence of robust domain specific system.

2.1 Introduction

In recent years, Natural Language Processing (NLP) has significantly advanced. Now machines can better understand, interpret and generate human language. From the many applications of NLP, Question-Answering system plays a very crucial role in search engines, customer support, virtual assistant and educational platforms [5]. While QA research and development have spread in high resource languages like English and Chinese, low resource language like Bengali remain underexplored [6]. Even Bengali is the seventh most spoken language globally.

To address this gap, this research proposes the creation of a Bengali QA dataset which is in jsonl format. The goal is to provide a clean, diverse and publicly available resource which can be used to train and evaluate QA models for Bengali.

2.2 Literature Review

Extensive research has been conducted in Question-answering system for high-resource languages. Several methodologies have been explored, ranging from rule-based systems and syntactic transformations to advanced deep learning models, especially transformer architectures like Multilingual T5, and GPT variants.Tn term of Bangla, the research is very recent and limited. Different studies have sought to develop Bangla QA systems by fine-tuning pre-trained models, building bilingual models. A few notable works include:

F. Rabbi and M. R. Biswas [1], the authors present a comprehensive framework for Bangla NLP .They have introduced an annotated corpus and implemented some key tasks like tokenization, POS tagging, named entity recognition, and sentiment analysis using Python-based machine learning approaches.Their project have an accuracy of 93.94%, this project

provides a strong foundation for advancing BNLP, specially by offering open-source resources for future research.

S. Shammi [2] focuses on sentiment analysis in Bangla. This work state of transformer-based approaches. It highlights the challenges of resource scarcity, complex morphology, orthographic variations, idiomatic expressions, and the lack of standardized NLP tools. The paper focuses on the necessity of developing culturally sensitive datasets and methodologies for advancing sentiment analysis in Bangla.

In [3], a domain-specific question answering (QA) system is proposed using a fine-tuned BERT-Bangla model. The system is trained on 2,500 question–answer pairs from Khulna University of Engineering & Technology (KUET) and achieves an Exact Match score of 55.26% and an F1 score of 74.21%. The results are promising in this project but the work shows limitations in handling complex or ambiguous queries and underscoring the need for further refinement.

A hybrid sentiment analysis approach is introduced in [4]. It combines a novel rule-based algorithm, Bangla Sentiment Polarity Score (BSPS), with BanglaBERT. The model categorizes text into nine sentiment classes, moving beyond binary or ternary sentiment analysis. The hybrid method outperforms standalone BanglaBERT, demonstrating the effectiveness of integrating rule-based and transformer models for nuanced sentiment analysis in Bangla.

T. Bhattacharjee,[5] presents BanglaBERT, a large-scale BERT-based language model pretrained on 27.5 GB of Bangla text (Bangla2B+).Also proposed the Bangla Language Understanding Benchmark (BLUB), where sentiment classification, natural language inference, named entity recognition, and question answering are covered. BanglaBERT significantly outperforms multilingual models (mBERT and XLM-R), marking a milestone in BNLP by providing robust benchmarks and resources for future research.

These studies illustrate the steady growth of BNLP research, moving from resource creation and corpus development toward sophisticated applications like sentiment analysis and question answering.The field still have challenges like limitation of datasets, domain-specific adaptability, and the need for scalable tools. Addressing these gaps will be critical for establishing Bangla NLP at par with resource-rich languages.

Table 2.1: Summary of Literature Reviewed.

Author (s)	Year	Title	Methodology	Key Findings
Rabbi, F. & Biswas, M.R.	2024	Annotated Bangla Natural Language Processing (BNLP) Using Python and Machine Learning	Tutorial-style methodological approach introducing annotated BNLP tasks using Python ML libraries	Demonstrates practical implementation of NLP tasks Bangla with Python; highlights importance of labeled data and language specific tools
Sengupta, S., Ghosh, S., Mitra, P., & Tamiti, T.I.	2024	Milestones in Bengali Sentiment Analysis leveraging Transformer-models: Fundamentals, Challenges and Future Directions	Literature review and comparative survey of transformer-based sentiment analysis systems in Bengali	Transformer-based models significantly outperform classical ML approaches; identifies major challenges due to data scarcity and linguistics complexity
Roy S.C. & Manik, M.M.H.	2024	Question-Answering System for Bangla: Fine-tuning BERT-Bangla for a Closed Domain	Fine-Tuning Bangla-BERT on a domain-specific QA dataset	Closed-domain fine-tuning improves QA performance; Bangla-BERT can be effective adapted for Bangla question answering with small dataset
Mahmud, H., Mahmud, h. & Rashid, M.R.A.	2024	Enhancing Sentiment Analysis in Bengali Texts: Hybrid Approach Using Lexicon-Based Algorithm and Pretrained Language Model Bangla-BERT	Hybrid model combining lexicon-based sentiment scoring with Bangla-BERT embeddings	Hybrid approach yields higher sentiment classification accuracy compared to standalone lexicon or transformer models
Brattacharjee, A. Et al.	2021	BanglaBERT: Language model pre-training and benchmarks for low-resource language understanding evaluation in Bangla	Pretraining of a large transformers language model on Bengali corpora and evaluating on downstream benchmarks	BanglaBERT sets new state of the art on multiple Bangla NLP tasks, proving the benefit of large scale pretraining in low resource languages.

2.3 Similar Applications

Different QA systems is created for high-resource languages. Models like LLaMA, Gemma, Google's BERT and Meta's RoBERTa have great capabilities in multilingual QA tasks. However, these systems are have not full potential for Bengali language.

mBERT, XLM-RoBERTa, BanglaBERT and IndicBERT are some of the recently researched topic and projects for low-resource languages. There are few works attempted for Hindi and Chinese language but effort for Bengali is limited. LLaMA and BNLP tools show potential but require further Fine-tuning for better performance.

2.4 Gap Analysis

By researching and analysing some of the previous papers we have found some gaps and also included the contribution in this work.

Table 2.2: Gap Analysis & Contribution

Features	**Existing Datasets**	**Ideal Requirement/ Benchmark**	**Our Dataset**
Language Support	Mostly in English/Hindi	Bengali/	Bengali
Number of QA Pairs	Low	Good for training	e.g., 15000+
Domain Coverage	Limited	Diverse	General knowledge
Answer Type Diversity	Mostly conversational	Short+Long	Mostly factual
Data Format	Json, csv or plain text	Jsonl/structured schema	jsonl
Annotation Quality	Machine generated	Human verified	Manually verified
Context Availability	Often no context	With context (Paragraph/doc)	No context
Licensing & Accessibility	Few public, mostly private	Open-source/ accessible	Open-source

2.5 Summary

Natural Language Processing (NLP) is taking more attention. But low resource language like Bengali is underrepresented. Most available resources focus on English. While there are some Bengali datasets exist. They often lack of structure, consistency and rarely available in jsonl format which is likely more compatible with machine learning tasks. On the other hand these datasets frequently lack of manual validation, relevance or accuracy. This project addresses these gaps by creating a vast dataset, manually validated and properly structured Bengali QA dataset which can support research and development in Bengali language AI and NLP systems.

Chapter 3

Research Methodology

This chapter provides a comprehensive overview of the research methodology adopted for developing a Bengali QA system with the multilingual model LLaMA 3.1. It will cover the general design, system design, data flow, requirement specifications and project planning as per the corresponding engineering project.

3.1 Methodology

3.1.1 Overview

This chapter describes the methodology to develop a Bengali Question Answering (QA) dataset and fine tune a QA model using this dataset. The research was carried out in different phases including dataset design, data collection, formatting, validation and evaluation.The primary goal of this research to create a structured and diverse dataset of Bengali QA pairs and a QA system that can serve as a benchmark for Bengali NLP tasks.

3.1.2 Proposed Methodology

The proposed methodology describes the processes for developing the QA model. The methodology is divided into following phases:

i) Requirement Analysis: In this phase both the functional and non functional requirements have been identified. A context diagram created to define the scope.

ii) System Design: Data flow diagram and UI architecture were developed to define the workflow of the application.

iii) Implementation: The system was developed using Python, for interface Next.js and for API inference used Flask.

iv) Testing: Integration testing was performed to ensure the system's scalability and reliability.

v) Evaluation: The system's performance and responsiveness has been evaluated.

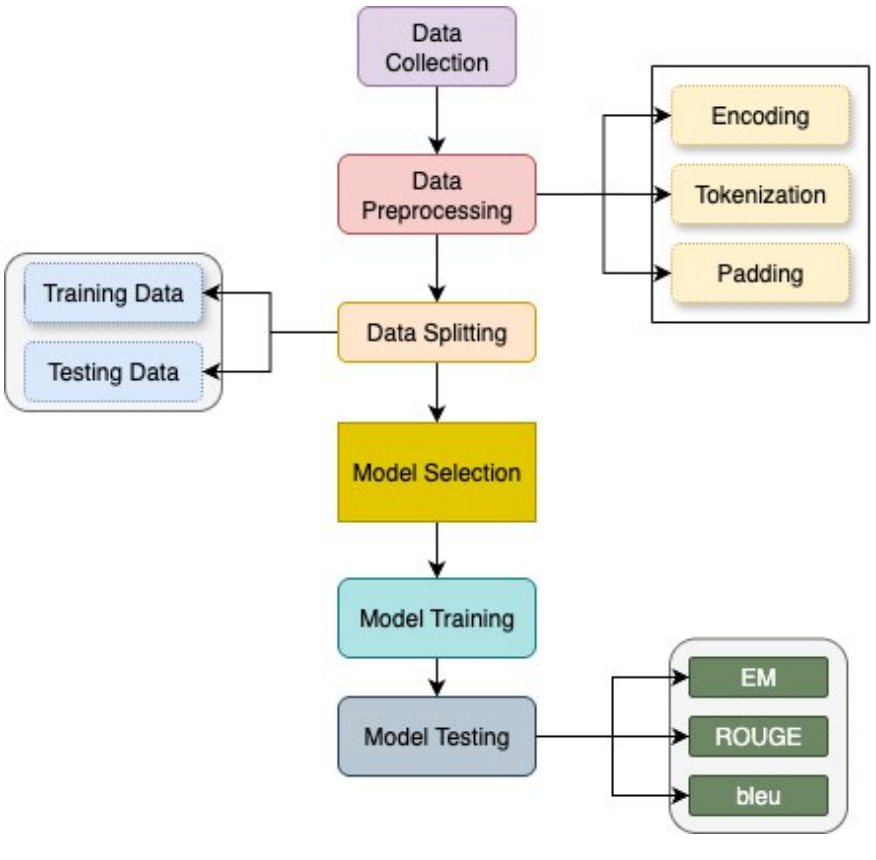


Figure 3.1: Methodology pipeline diagram

3.1.3 Functional and Non-Functional Requirements

A better understanding of the system's functional and non-functional requirements is crucial for guiding the design and development of the Bengali QA system.

Functional Requirements:

i) The system must accept user queries in Bengali.

ii) It must return correct answers based on the trained data.

iii) It must allow dataset updates for continuous improvement.

iv) The system must have a interface for the users.

v) The system should be available from anywhere.

vi) It should secure the personal data of a user.

Nonfunctional Requirements:

i) The system should respond within 1–2 seconds

ii) It must support Unicode Bengali characters.

iii) The interface should have both the light and dark theme feature.

iv) The system should be available at least 95% of the time.

v) The system should be able to handle many users at a time.

vi) The system should be scalable and secure for future integration.

3.1.4 Context Diagram

The context diagram illustrates the interaction among the data source, dataset, annotator and the researcher. Its give us a simplified overview of building the primary dataset.

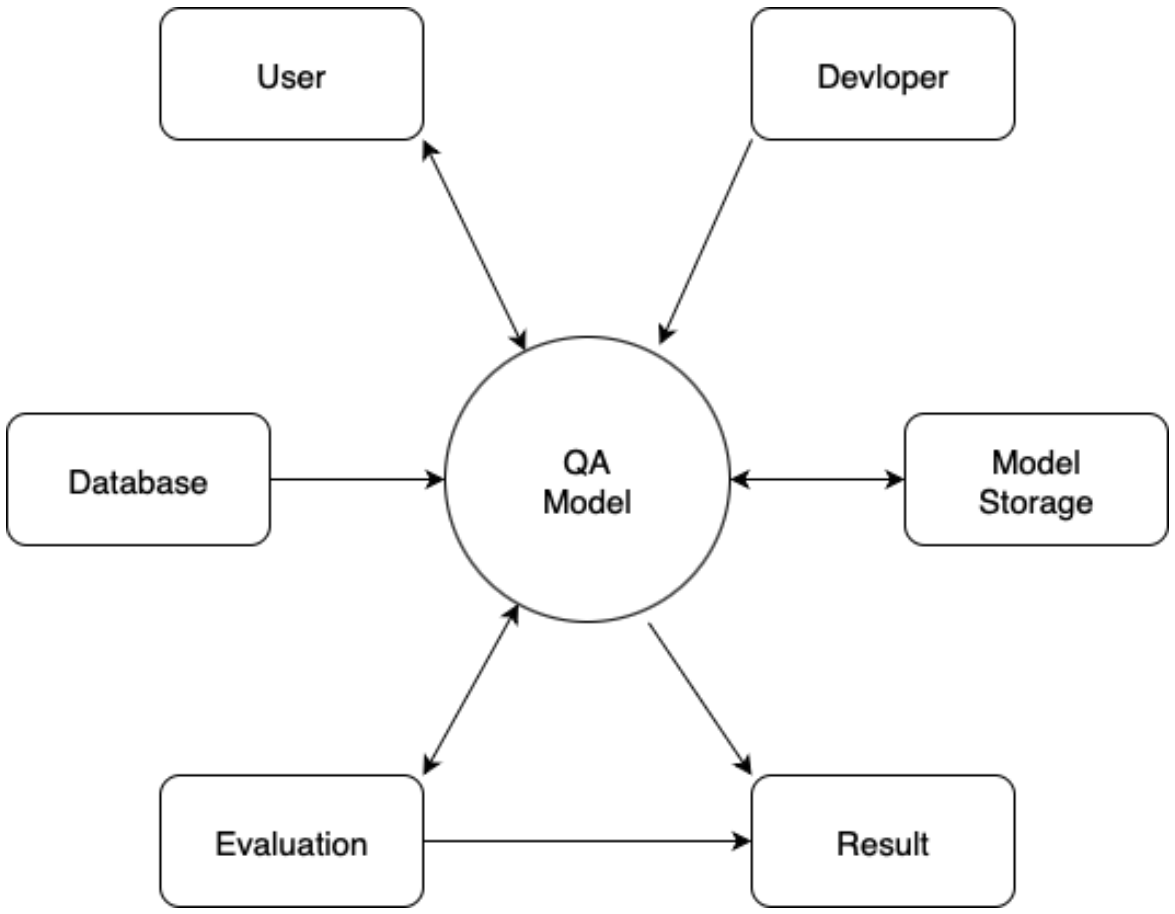


Figure 3.2: Context diagram of the system

3.1.5 Data Flow Diagram

The Data Flow Diagram (DFD) represents the flow of data within the system. It defines how data is collected, created, processed and output through the system. DFD's are used to visualize the logic and processes of the system. Here is added a DFD diagram for better understanding.

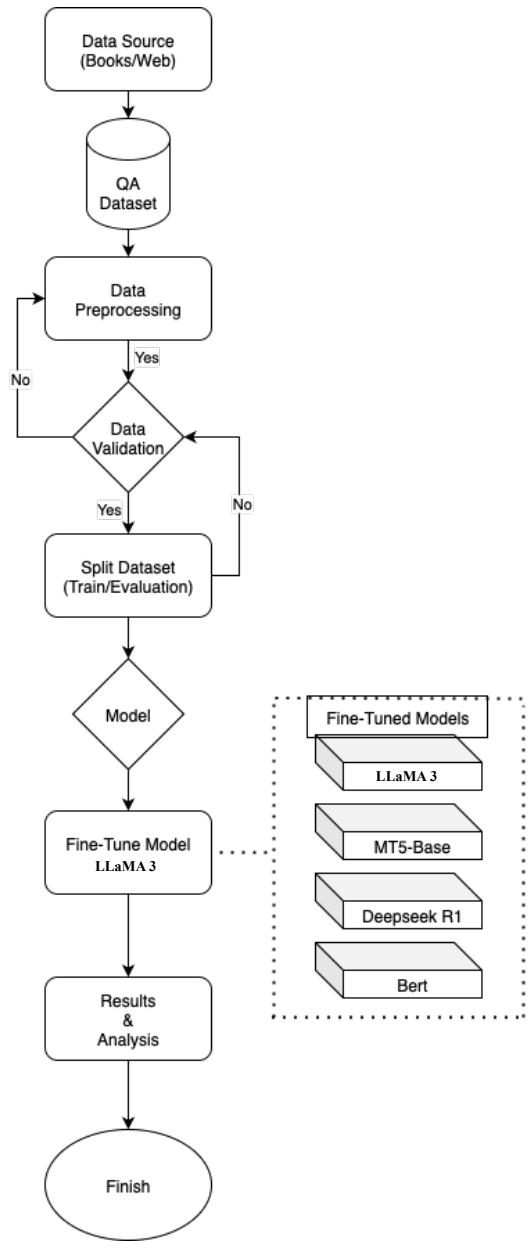


Figure 3.3: Data Flow Diagram

3.1.6 UI Design

Our primary goal is to Develop an efficient question-answering system for Bengali language by fine-tuning a pre-trained model, significant attention was also given to designing and implementing a user-friendly graphical user interface (GUI).The result is a complete and interactive web-based system that enables teachers, learners, and researchers particularly those without a technical background to easily access and utilize automated Bangla question generation.

The current interface offers an intuitive and responsive platform where the user can input Bangla text as question and instantly receive answer of the question. Key features of the UI includes an input box, realtime answer visualization, user can adjust the brightness for day-night mode, the previous saved chat (if logged in), a new-chat option. The frontend has been designed using Next.js.

To ensure accessibility and usability across a wide range of devices, the interface follows responsive design principles and prioritizes user experience. While the main functionalities are implemented, the system is structured for future enhancements such as batch uploads, authentication, analytics, and downloadable outputs, making it scalable for classroom, individual, or research use.

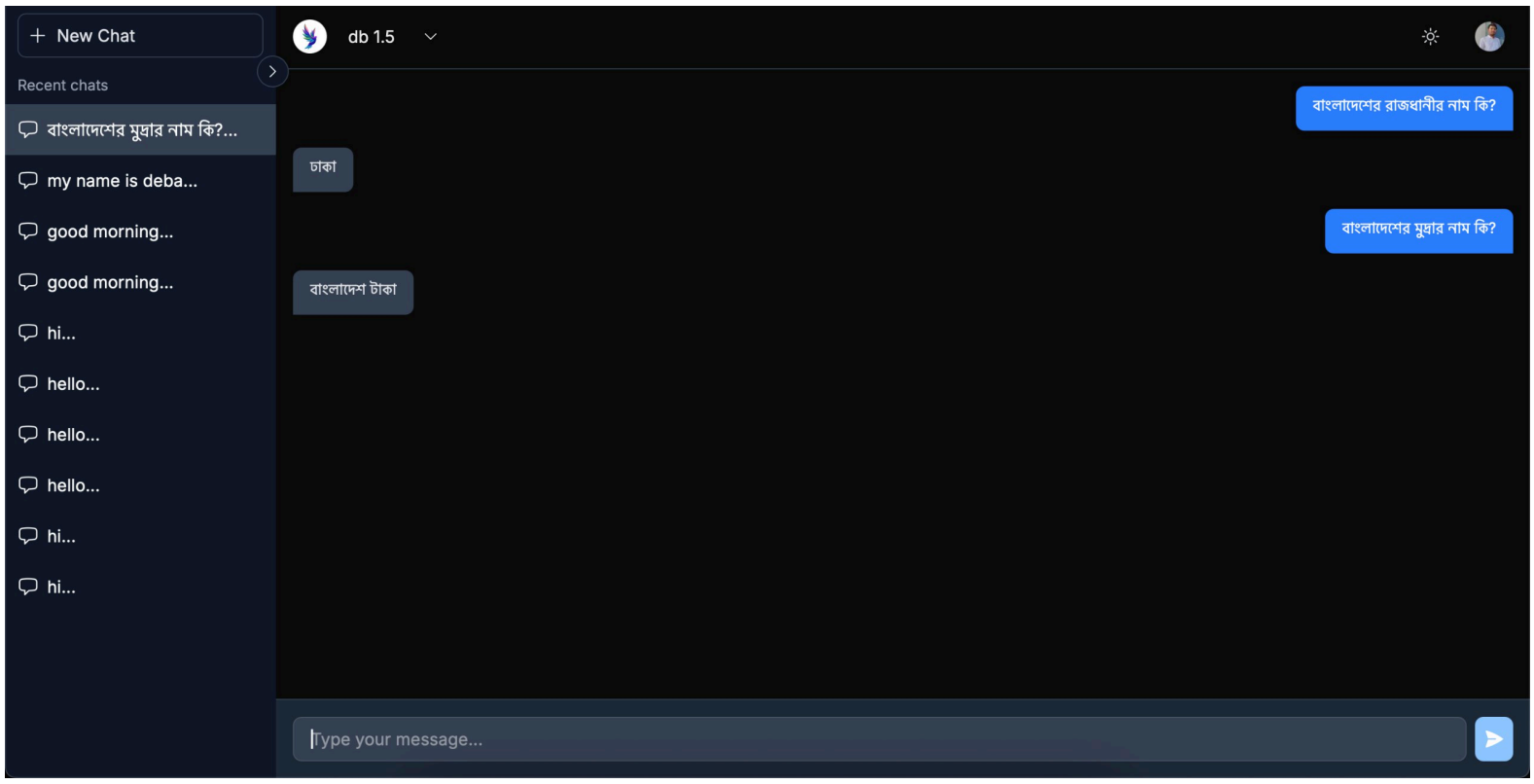


Figure 3.4: UI design

3.2 Detailed Methodology and Design

This research focused on building a structured Bengali QA dataset. The research follows the following stages -

3.2.1 Data Collection:

The data collection phase was very time consuming and a hard working process. The dataset is made manually and collected the data from various types of general knowledge books such as Ajker Bissho, Bissho Bichitra and searched in the web for various information. Different types of domain is selected for the dataset.

3.2.2 Question Answering Pair Design:

Each item structured as - {“input”: “বাংলাদেশের রাজধানী কি?”, “output”: “ঢাকা”}. We used here mostly the fact based questions like what, who, when, where, etc. Answers are short, direct and in Bengali.

3.2.3 Preprocessing:

Preprocessing technique is very crucial before training models. Each data is pointed to QA pair, which presented as input-output generation task. The preprocessing has started with text normalization which means punctuation marks, special characters has been normalized. Unnecessary white spaces removed and the most important Unicode normalization has been done to use the encoding techniques.

We have implemented the custom AutoTokenizer for tokenizing the encoded data. We have customized the tokenizer with chat template. For model LLaMA 3.1 we also utilized the SentencePiece tokenizer. Each question and answer has been tokenized to their max length. For the small length question and answer we used the pad token to fill the empty tokens.

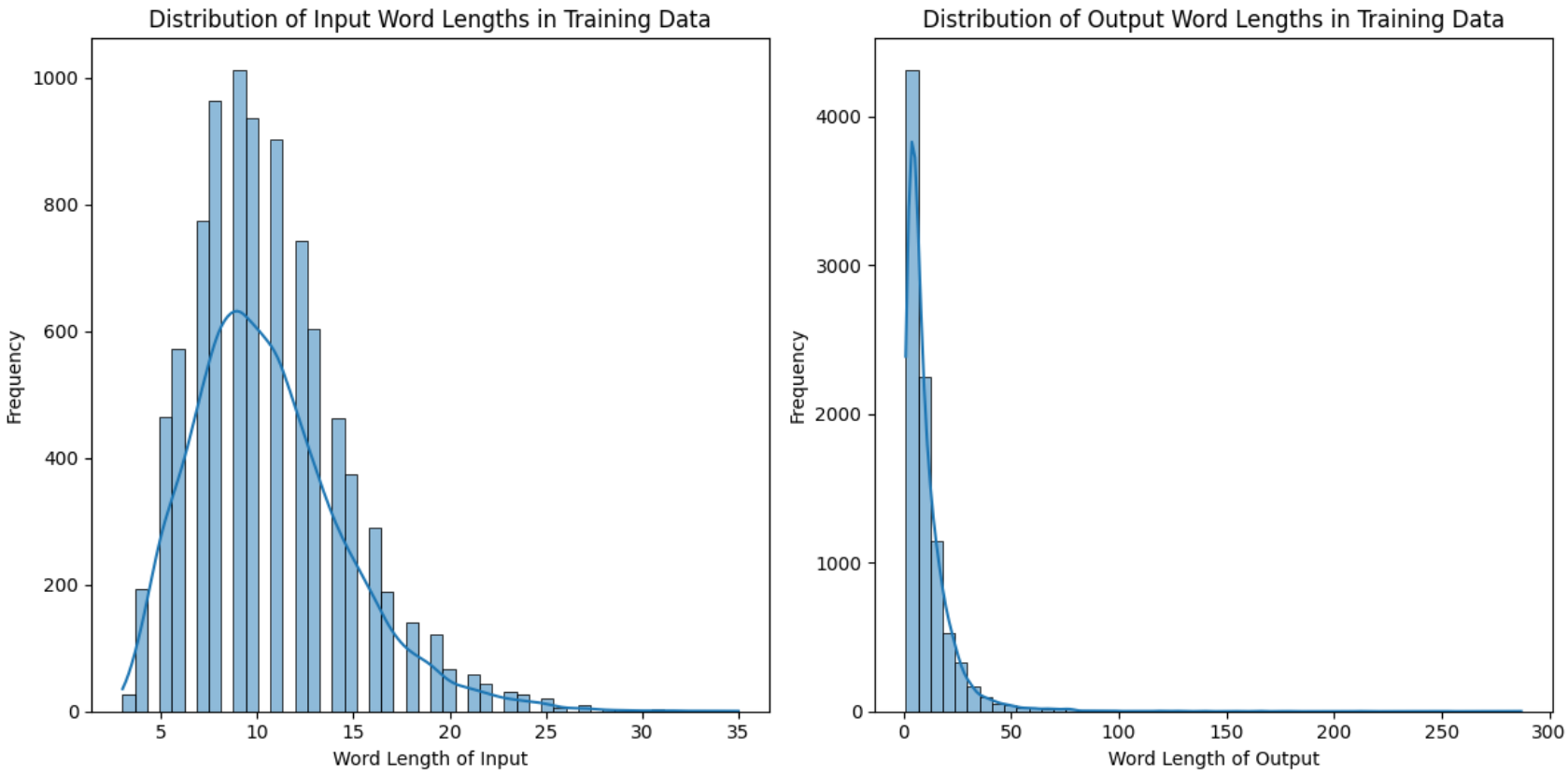


Figure 3.5: Word length distribution of Training data

Figure 3.5. "Word Length Distribution of Training dataset" presents two histograms, distribution of input word lengths and distribution of output word lengths. In the left plot, there are high frequency words in the input data. On the right side of the plot, training output data word length is comparatively less than the input data.

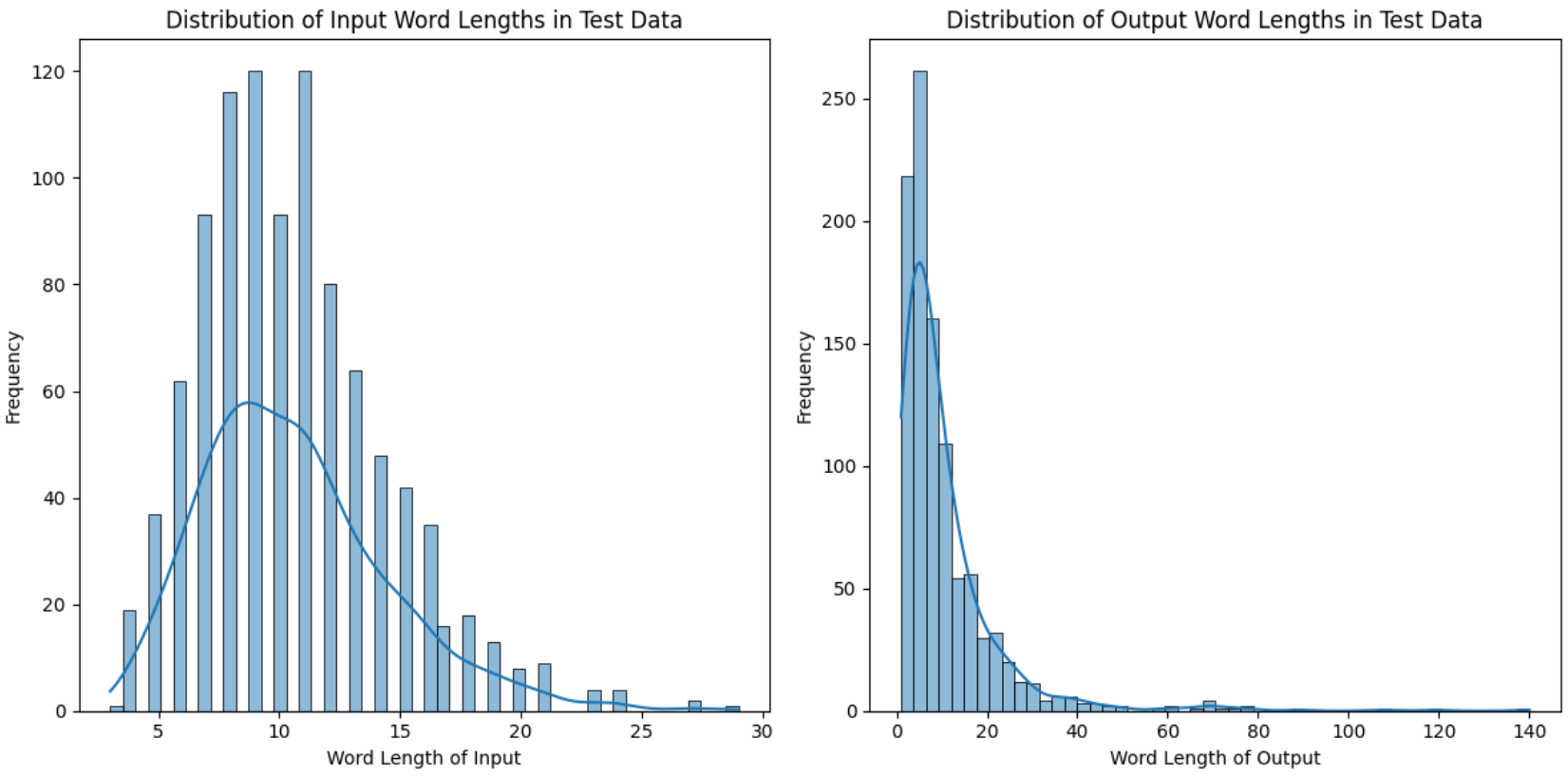


Figure 3.6: Word length distribution of Test data

In the Figure 3.6 “Word length distribution of Test data” presents two histograms, distribution of input word lengths and distribution of output word lengths. The left plot of the histogram shows that the word length and frequency of input data. On the other hand right plot of the histogram shows the word length and frequency of output data which comparatively less than the input data. This indicates that answers are shorter than the questions which will lead to the consistent model training.

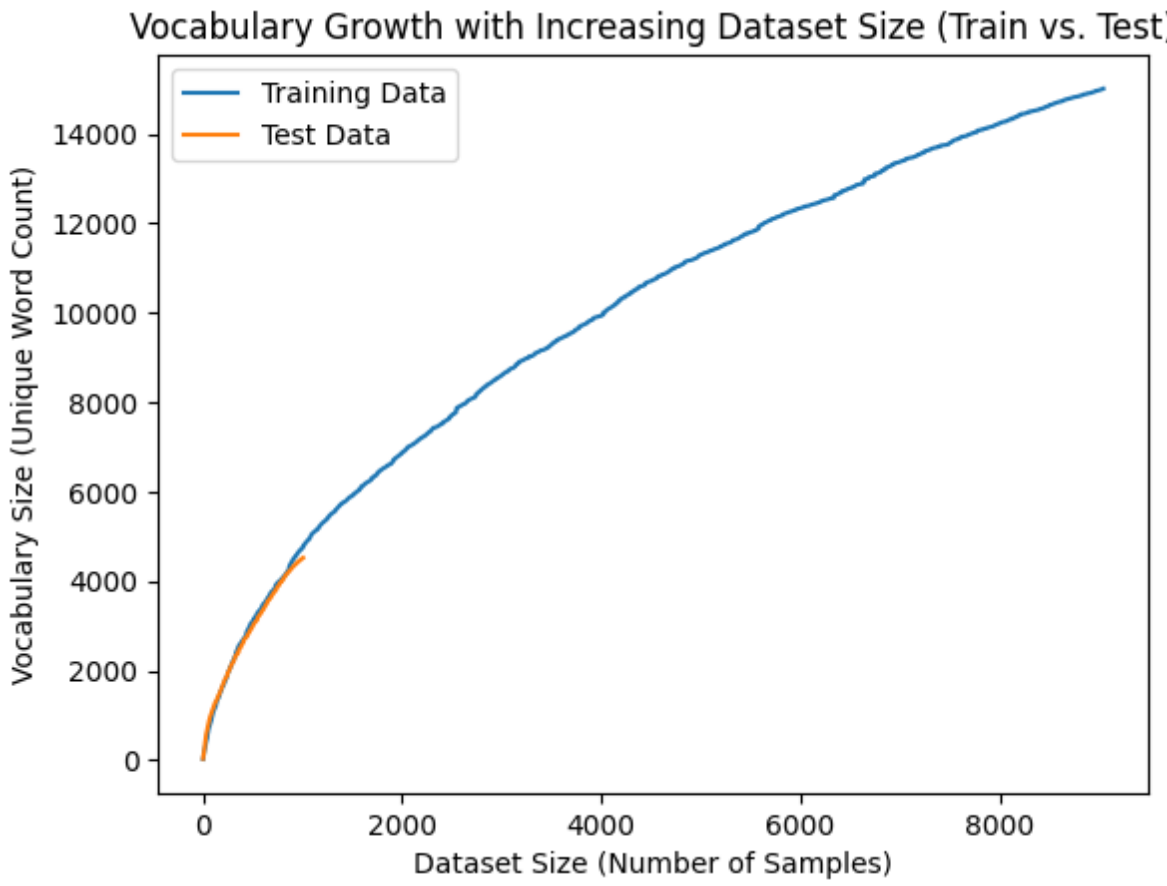


Figure 3.7: Vocabulary Growth of dataset

Figure 3.7 titled “Vocabulary Growth of dataset” indicates the unique word count of the dataset. Based on the dataset the training data has more than 14000+ unique words and in the test data the are about 4500 unique words. Thus its produce a strong vocabulary in the model.

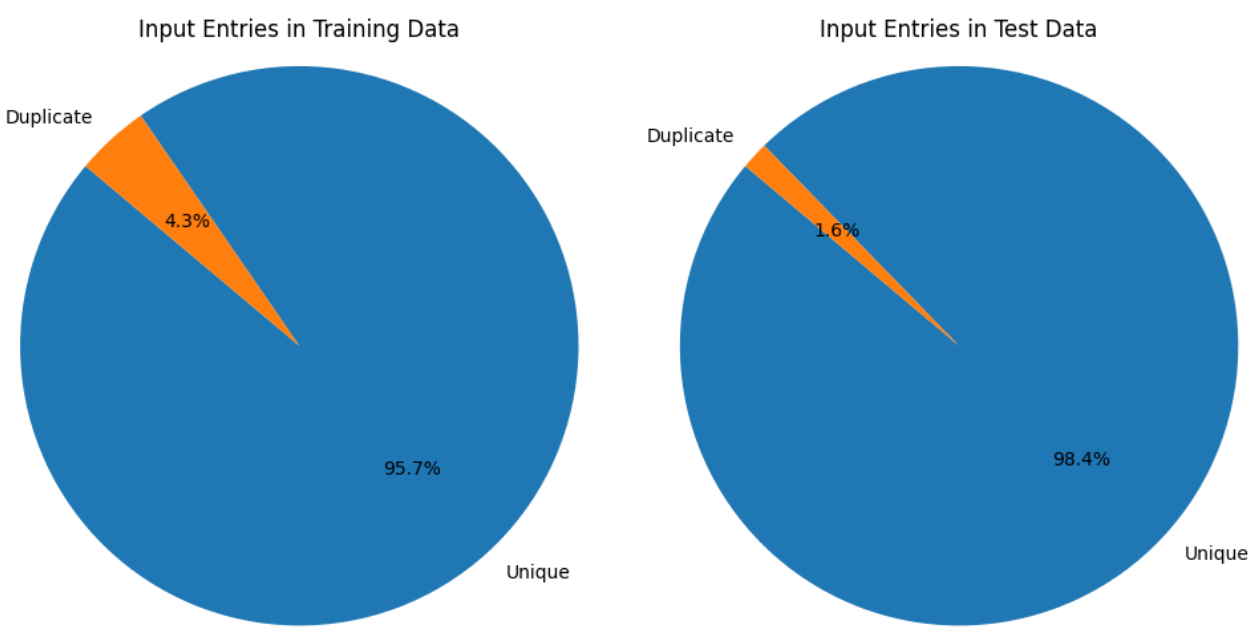


Figure 3.8: Duplicate value count

By the PyChart in Figure 3.8 “Duplicate value count” 95.7% of data are unique and 4.3% data is duplicate only on training data. Beside on test data 98.4% data is unique and only 1.6% of data is duplicated. These uniqueness will help the model to learn and evaluate to enhance performance.

3.2.4 Model Fine-Tuning

The setup for fine-tuning model LLaMA 3.1 was totally different. Its a large language model, supported Bengali and other languages. To train the model we have used SFTTrainer with DataCollatorForSeq2Seq. Training_args used the SFTConfig with 50 epoch, 1e-4 learning rate and adamw_8bit optimizer. Per device train batch size was 2, gradient accumulation steps was 4. The model was trained with train on response only. After every epoch finished LLaMA 3.1 performed evaluation to evaluate the performance of the model and automatically save the best checkpoint into the output directory. Training and evaluation was performed using the A5000 GPU. After training successfully the model adopted to perform Bengali Question-Answering with high accuracy.

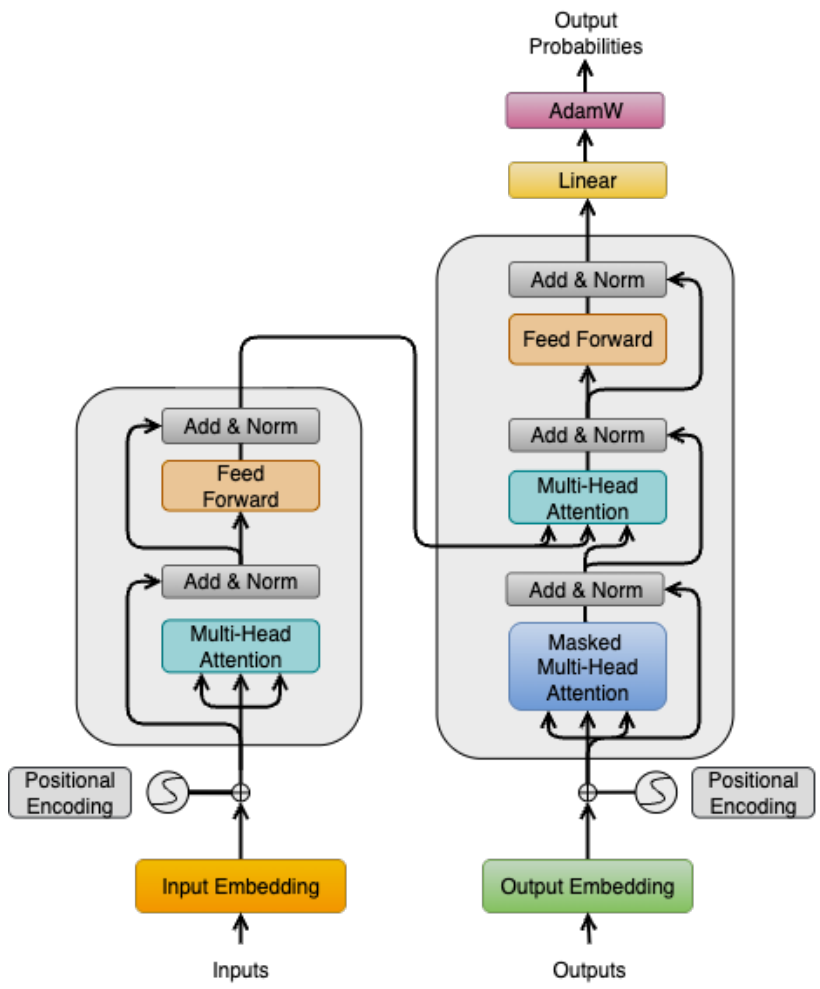


Figure 3.9: Architecture of LLM

3.2.5 Evaluation and Metrics

After finished the fine-tuning process an evaluation was performed to evaluate the performance of the model LLaMA 3.1. The metrics are EM, BLEU and ROUGE. These metrics are specially designed to measure the correctness and relevancy with the prediction and the references.

A special library called evaluate was implemented to measure the EM, BLEU and ROUGE score. The Exact Match (EM) metric is very strict. Its always search for the exact match with the references. BLEU metric is quite less strict. It measures the n-gram precision means how much of prediction matches the references. ROUGE metric is flexible, it measures the n-gram recall and overlap means how much of references is covered. Implementation of the metrics are in the below -

$$Precision = \frac{LCS}{candidate\ length}$$

Here LCS is the longest common subsequences between the candidate and references.

$$Recall = \frac{LCS}{reference\ length}$$

$$EM = \frac{Number\ of\ Exact\ Match}{Total\ number\ of\ samples}$$

$$ROUGE\text{-}N = \frac{\sum_{ngram \in \text{Reference}} \min\left(\text{Count}_{candidate}(ngram), \text{Count}_{reference}(ngram)\right)}{\sum_{ngram \in \text{Reference}} \text{Count}_{reference}(ngram)}$$

$$R_{LCS} = \frac{LCS(X,Y)}{|Y|}, \quad P_{LCS} = \frac{LCS(X,Y)}{|X|}, \quad ROUGE\text{-}L = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}}$$

Typically, these metrics can provide a comprehensive performance in model performance. BLEU highlights precision and n-gram overlap, EM measures exact correctness and ROUGE highlights recall. Using these three metrics in combination ensures a reliable assessment in NLP models.

3.2.6 Deployment

To illustrate the model's assumption we have created a frontend using Next.js and Flask as backend to perform the overall combination.

I. Load Model: The fine-tuned model is hosted in the https://huggingface.co/spaces/thedeba/debai
II. API Integration: Hosted model's API is used in frontend to generate responses from the model.
III. Input Field: A input text field is added to enter Questions.
IV. Result: After getting the input, generated answer will be displayed on the UI.

3.3 Project Plan

After considering the complexity of the project, project planning was made very precisely. It has also been ensured that Bengali QA system was developed, tested and deployed. For smooth delivery in time we divided the project into several phases-

Phase 1: Research and Requirement Analysis

i. Started through the literature review on QA system in Bengali.

ii. Studied documentation on multilingual models that support Bengali.

iii. Focused on text generation to maintain clear evaluation.

Phase 2: Dataset Curation and Preparation

i. Started creating QA pairs to create a structured dataset.

ii. Performed multiple preprocessing and cleaning process to the dataset.

iii. Dataset is designed to compatible with HuggingFace Transformers format.

Phase 3: Model Fine-Tuning

i. Developed training script for training the pre-trained model.

ii. Platform jarvislabs ai selected for efficient training.

iii. Conducted fine-tuning with several hyperparameters on the dataset.

Phase 4: Evaluation and Testing

i. Used BLEU, ROUGE and EM metrics for evaluation.

ii. Compared with Gemma 3, mT5-base, LLaMA 3.1 models performance.

iii. Depending on the test result selected the LLaMA 3.1 model.

Phase 5: Deployment

i. For the UI design used the Next.js for Server Side Rendering (SSR) and scalability.

ii. API handling on the backend was implemented using Python Flask library.

Tools and Technology used:

i. Programming Language: Python 3.11

ii. Deep Learning Framework: PyTorch, Transformers, HuggingFace

iii. Environment: jarvislabs ai

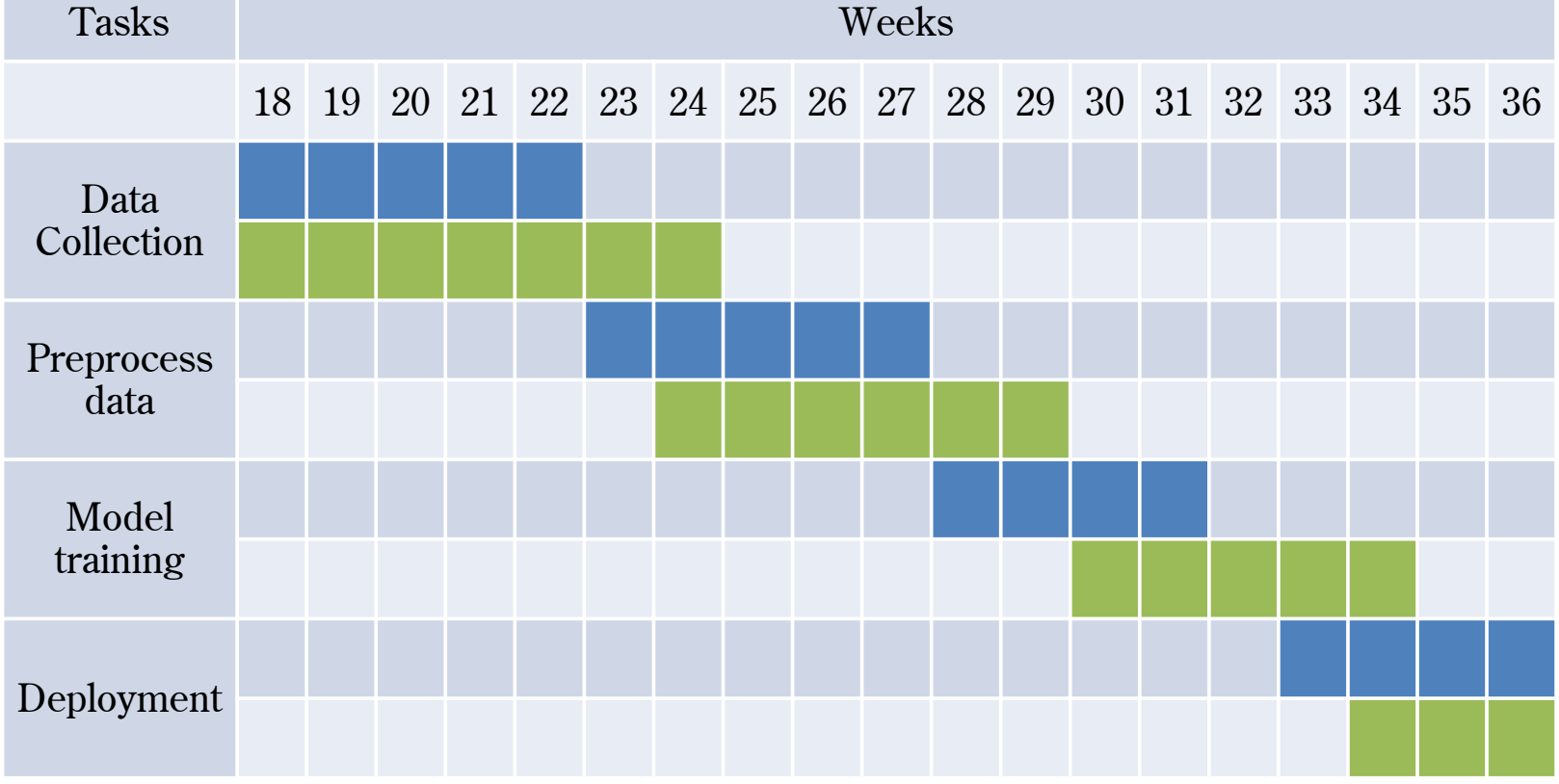
iv. Evaluation Library: evaluate, BLEU, ROUGE score

v. Deployment: Next.js, Flask

3.4 Task Allocation

Data collection phase was heavily time consuming, it has taken nearly 7 weeks to complete. After completing the data collection we have started preprocessing the data in week 24 to week 29. The real timeline exceeded the expected timeline of the project. Model training phase started in week 30 and finished in week 34. After getting all things ready started deployment in week 34 .

Table 3.1: Task Allocation Table



The Table 3.1 shows the timeline of each task of the project from week 18 to week 36. It starts from the data collection in week 18. Sequentially after data collection, preprocess data, model training and the deployment phase.

3.5 Summary

This research involved the systematic development of a high quality Bengali Question-Answer (QA) dataset and a QA system. The process began with the manual collection of data from the textbooks and online Bengali sources. Short questions and fact based questions answers are stored in the dataset. We have reviewed the data and grammatical correctness. The validated data was stored in a standardized .jsonl format. After collecting data several preprocessing steps were performed. Fine-tuning LLaMA 3.1 with the dataset achieved a great performance. A user friendly UI has integrated to ensure the deployment.

Chapter 4

Implementation and Results

This chapter gives a description of real world implementation of the Bengali QA system using the LLaMA 3.1 model. It provides information about environmental setup, dataset preparation, training procedures, testing model and methods of evaluation. It also discusses about the outcome achieved during the development.

Every implementation was performed in a structured manner to achieve the best training, efficient deployment and moderate output of question answering.

4.1 Environment Setup

Before developing the model we have to set up the environment. So we have worked with 2 (two) types of environment. They are -

4.1.1 Hardware Environment

To train the model the project need a significant computational resources for its size of the LLaMA-3 model and the curated Bengali QA dataset. The training process was carried out using high-performance GPUs such as NVIDIA A6000, RTX6000 Ada, RTX5000, and A5000, each offering between 12–24 GB of VRAM.

The system interface was designed to be more flexible so that out can run on both CPU or GPU.The memory storage requirements are at least 16 GB RAM to handle large batch sizes and dataset loading. For interface 8 GB RAM is sufficient. Storage requirement is approximate to 100 GB to accommodate the model, dataset, and checkpoint files during training while 25 GB was adequate for interface purpose.

The setup can successfully run on Ubuntu, Windows, and MacOS, which ensures cross-platform compatibility.

Table 4.1: Hardware Environment

Component	Training	Inferences
GPU	A6000, RTX6000Ada, RTX5000, A5000 (12-24GB VRAM)	Can run CPU or GPU
RAM	16GB+	8GB+
Storage	~100GB (model+dataset+checkpoints)	25GB
OS	Ubuntu/Windows/Mac	Ubuntu/Windows/Mac

4.1.2 Software Environment

The model was trained using Python 3.9 and 3.10, which are widely supported by modern deep learning frameworks.These ensure the compatibility and stability across all stages of development.

The basic deep learning framework used was PyTorch 2.1.0+, this is really flexible and robust that support for transformer-based architecture. For GPU acceleration, CUDA versions 11.8 or 12.x is used, that optimized the computation.

Several specialized libraries were integrated into the environment:

i. Transformers (v4.39) – for model loading, fine-tuning, and integration with Hugging Face APIs.

ii. Datasets (v2.18) – for efficient dataset handling, preprocessing, and batching.

iii. Accelerate (v0.26) – for seamless multi-GPU and distributed training.

iv. PEFT (v0.7) – for parameter-efficient fine-tuning, which reduces computational overhead while maintaining performance.

v. Evaluate (≥0.4.0) – for model performance assessment using metrics such as ROUGE and BLEU.

Table 4.2: Software Environment-1

Library	Recommended Version
Python	3.9 or 3.10
PyTorch	2.1.0+
CUDA	11.8 or 12.x
transformers	4.39
datasets	2.18
accelerate	0.26
peft	0.7
evaluate	>=0.4.0

4.2 Test Performance Evaluation and Comparative Analysis

The Bengali QA system was evaluated using the test set of question answering pairs which was early splitted from the dataset. Fine-tuned model was loaded in the environment with PyTorch. For each question from the test dataset the model generated an answer which was compared with the references to get the evaluation result.

Automated evaluation was conducted using standard evaluation metrics named BLEU and ROUGE. BLEU measures the n-gram overlap in the generated output and the references output of similarity both in syntax and lexicon. LLaMA 3.1 obtained at 0.5346. Concurrently, ROUGE score focus on recall based matching at the level of phases. ROUGE-L measures the longest common subsequences in candidate versus reference output. The model LLaMA 3.1 obtained 76.38%. This discovery highlights the supremacy of language specific pre-training in LLaMA.

Table 4.3: Comparative Model Performance

Authors	Dataset Size	Model	Result
Saptarshi Sengupta	158065	BanglaBERT	63%
Subal Chandra Roy	2500	BanglaBERT	55.26%
Hemal Mahmud*	15194	BanglaBERT	74%
Abhik Bhattacharjee	5.25M	BanglaBERT	77.09%
Proposed	15000	LLaMA	72.37%

A comprehensive evaluation of the model BanglaBERT performance in various research works and the proposed system is given in Table 4.3, where outputted F1 score of all the

previous research are noted. Because of BanglaBERT is a non-multilingual model it only works with Bengali. So our proposed model is multilingual and showed a extraordinary performance with the raw dataset.

Saptarshi Sengupta [2] has implemented a fine tuned approach to the model BanglaBERT. Author used the dataset with 158065 examples. After fine-tuning the model in evaluation 63% F1 score is considered with the dataset.

Subar Chandra Roy [3] utilized the BanglaBERT model in the research. The dataset author used to fine-tune is quite small that only have 2500 examples. Because of using a small dataset the model is not gained a satisfactory performance. The overall F1 score is adopted is 55.26%.

Hemal Mahmud [4] developed a novel rule based algorithm Bangla sentiment polarity score with generating sentiment score. Author used to fine tune the model using 15194 examples. The BanglaBERT hybrid approach evaluated with 74% F1 score.

Abhik Bhattacharjee [5] introduced the BanglaBERT with two downstream task datasets on natural language inference and question answering and benchmark on four diverse NLU task covering text classification, sequence labelling and span prediction. They produced 5.25M examples of data with data crawling to train the model and gained 72.37% F1 score on the dataset.

In Comparison with other models, our proposed methodology did a outstanding performance with the raw dataset. Using fine-tuning method LLaMA as a base model, the system obtained a high ROUGE-L score of 76.38%. The findings underscore the effectiveness of using well annotated data and the language specific model.

4.3 Results and Discussion

For testing the model LLaMA, we used the test dataset's QA pairs and measured the standard evaluation such as BLEU, Exact Match and ROUGE (ROUGE-1, ROUGE-2, ROUGE-L). The goal was to ensure the output generation is grammatically correct, contextually appropriate and meaningful answers in Bengali.

Table 4.4: Performance analysis table

Metrics	Fine-Tuned Model
BLEU	0.5346
ROUGE 1	79.02%
ROUGE-2	61.47%

ROUGE-L	76.38%

We use two widely used evaluation metrics for the proposed system. The reason behind choosing these metrics because of their effectiveness of capturing the quality of generated output for both content overlap and linguistic fluency. The result of the base model and fine-tuned model are presented in the table 4.3.

The table clearly showing that the fine-tuned model outperforms the base model across both evaluation metrics.

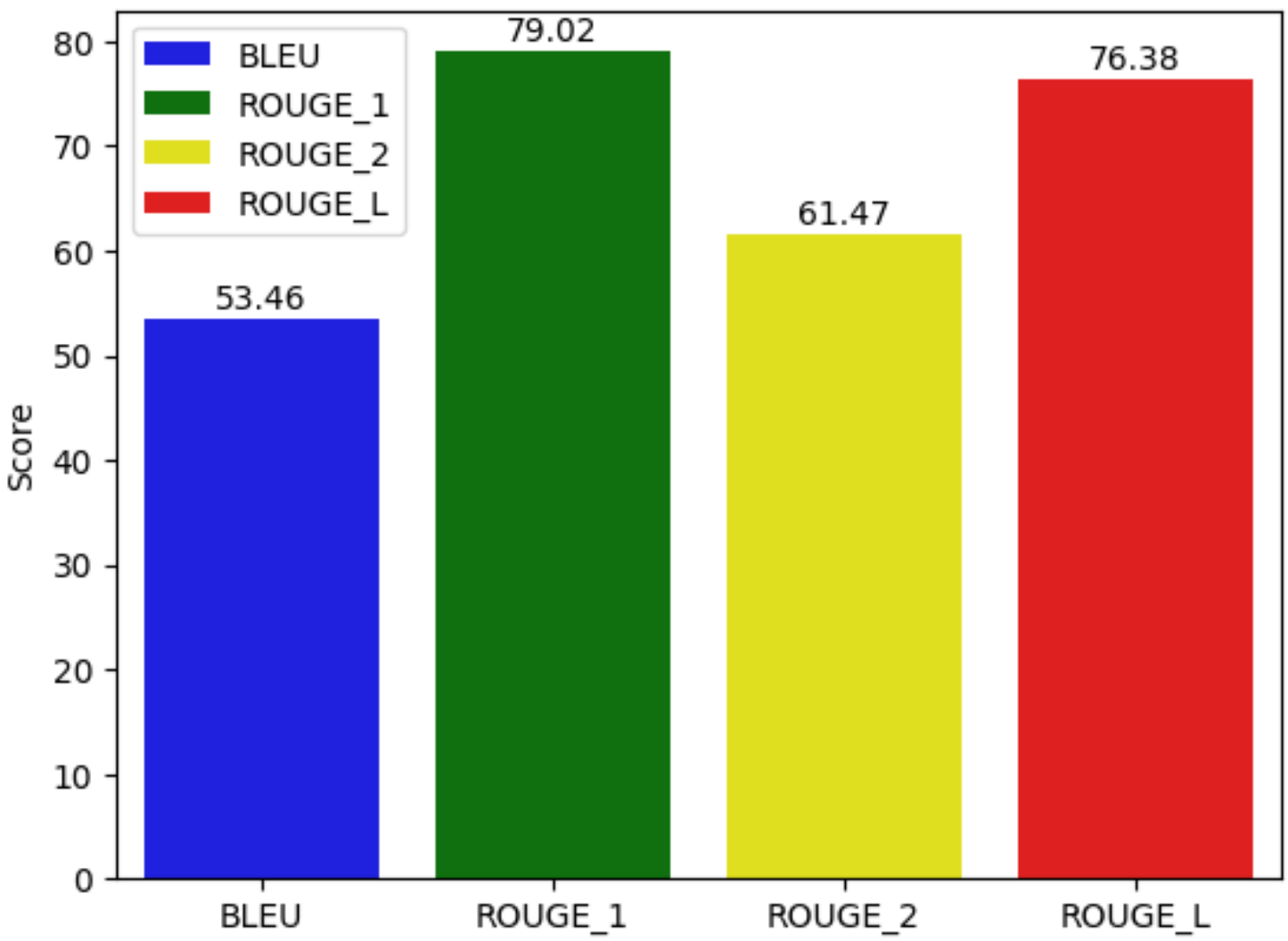


Figure 4.1: Performance Analysis

The ROUGE-L score of the base model was 0.48, that reflects a moderate level; of overlap between the generated text and the reference outputs.After fine-tuning, the ROGUE-L score increased to 0.57 what is a great improvement of 9 percent points.These improvement means that fine-tuning enhances the model's ability to generate responses that are pretty similar to the reference text.

The BLEU score is also 12 point improved in fine-tuned model then of the base model that means it increases fro 0.42 to 0.54. This improvement shows more fluency and accuracy of text sequences, with great precision in word choice and phrase construction. This score emphasizes exact n-gram matches and enhancing the quality of output which is generated.

These improvements in both metrics gives a clear positive impact of domain-specific fine-tuning.The fine-tune model performed better then of base model for equipping to capture contextual nuances, align with reference outputs, and generate coherent results. These enhancements are valuable for downstream applications like QA, summarization, or sentiment analysis in Bangla where semantic and linguistic accuracy are critical.

Table 4.5: Generated output of the model

Question	Original Answer	Generated Answer
কোন আইনের অধীনে রাজশাহী সিটি কর্পোরেশন গঠিত হয়েছে?	রাজশাহী সিটি কর্পোরেশন আইন, ১৯৮৭	রাজশাহী সিটি কর্পোরেশন আইন, ১৯৮৭
বিশ্বের সবচেয়ে বড় শহর কোনটি?	টোকিও	টোকিও
বিশ্বের সবচেয়ে বেশি বিক্রিত স্মার্টফোন ব্র্যান্ড কোনটি?	স্যামসাং	স্যামসাং
বাংলাদেশের মুদ্রার নাম কি?	টাকা	বাংলাদেশী টাকা

Table 4.5 provides few examples of some output generation of the fine-tune LLaMA model. It is seen that the original answer and the generated is mostly accurate.

Despite of the slight improvement the limitations is that the scored is not close to optimal, so there is still space for improvement. Factors like dataset size, diversity and quality of data that influence the performance of model. Future work may have larger dataset or advanced fine-tuning strategies and experiment with hybrid approaches that will combine rule-based and deep learning methods.

Fine-tuning a base model with a QA dataset improves the answer quality. The result confirms that specific language adaption is very crucial for underrepresented language like Bengali.

4.4 Summary

In this chapter, we explained the whole procedure of the implementation of the Bengali QA system project from environment setup to dataset preparation, model fine-tuning, testing and evaluation. The system was able to produce nearly fluent accurate answer for the questions. As a multilingual model it was slightly better than the other multilingual model.

Chapter 5

Engineering Standards and Design Challenges

In this chapter, we introduced the engineering paradigms, standards and some design challenges that we have faced while designing and developing the Bengali QA system. These include fidelity to software and hardware standards, communication protocol among the team, environmental and ethical consequences of the project. This research also compares the system to complex engineering tasks and problem solving tier.

5.1 Compliance with the Standards

In developing the Bengali QA system, compliance with software, hardware was considered to maintain reliability and sustainability of the project.

5.1.1 Software Standards

The following Bengali QA system is developed using standard software engineering and deep learning manner to ensure future extensibility, maintainability and ethical reproducibility. All source code comply with Python PEP 8 coding standards, clarity, modularity and consistency.

In terms of failure, exceptional exception handling has been integrated to ensure adaptability during model inference and data processing. The pre-trained model that have been used in this project from a trusted and valid platform called HuggingFace. After training the model saved it into the HuggingFace hub. From this platform we also hosted the model from the spaces and make an API implementation using Flask. All changes can be made with version control using git commands to uphold data privacy.

5.1.2 Hardware Standards

The system was run in a cloud based workstation called https://jarvislabs.ai/ to enhance the training. This cloud platform provides various types of GPU, CPU and storage. We have considered RTX5000 (7CPUs | 32GB RAM | 16GB VRAM), A5000 (7CPUs | 32GB RAM | 24GB VRAM), A6000 (7CPUs | 32GB RAM | 48GB VRAM), different types of instances for training, testing and evaluating. We also have used the gradient checkpointing and LoRA (Low-Rank Adaptation) techniques to reduce GPU memory usages.

5.1.3 Communication Standards

Although it's a research based project, standard network protocols has been followed for downloading the model and pushing it into the HuggingFace Hub. We used the JSON structure standard data format for question-answering pairs. JSON was selected because its lightweight and wide library support in Python.

We had conducted weekly meetings to reach our goal, discussed about problems and documented the results. Collaborative utilities made the development continued. These communication practices remain in place until deployment, maintainability, accountability and documentation driven updates of the insights.

5.2 Impact on Society, Environment and Sustainability

Under the development of the Bengali QA system to contribute meaningful impact including but not bounded to education, society, environment and ethical AI research.

5.2.1 Impact on Life

The Bengali QA system based on the Bengali language and help native users to retrieve accurate answers to their queries. This model can be helpful in future applications in education, research and digital literacy. It will also improve the access to knowledge for Bengali speakers.

The deployment plays an important role in increasing digital literacy and liability to AI tools within the Bengali speaking populations. The system is designed to be lightweight, particularly ideal for rural and remote areas where people have low bandwidth and has potential to support impartial access to educational technology.

5.2.2 Impact on Society & Environment

As a research project the social and environmental impacts are limited. The more bigger model we choose the more GPU is required in this project. As much as we use the GPU a heavy energy consumption can affect in the environment. So we have chosen a lightweight model to reduce the GPU usages.

Creating learning materials digitally reduces for printed copied and help to save valuable resources. At the same time, as the system is inclusive and open to everyone it gives poor communities the chance to access digital learning. In this way learners can overcome not only the technological barriers but also the challenges in learning experiences.

5.2.3 Ethical Aspects

Ethical aspects includes ensuring that the model does not generate harmful or biased

responses. The data is used for fine-tuning the model was carefully curated. As it is a research based project, ethical standards such as fairness, principles, responsibility and respect were followed for future development.

The system has made available for educational and non-commercial purposes. In its final deployment, accessibility was prioritized through interface design and localization. It also ensures that more people can benefit from the system.

5.2.4 Sustainability Plan

The sustainability plan means reproducibility and maintainability. By documenting all the codes, environmental setup, training configurations so that the system can be updated without starting from the scratch.

A long-term sustainability roadmap foundation deployment and increasing of the system. The dataset will be periodically updated with new domain specific content. Open sourcing the system encouraged academic border and community contributions.

5.3 Project Management and Financial Analysis

The project management and financial analysis defines the cost during the development of the project. To reduce the cost of hardware we used cloud based GPU like RTX5000, A5000 and A6000. Software costs were neutralised due to use of the open-source frameworks like PyTorch, HuggingFace libraries.

A detailed financial breakdown is given below:

Table 5.1: Estimated Cost and Financial Analysis

Components	Estimated Cost (BDT)
Cloud Computing (https://jarvislabs.ai)	2500-3000
Data collection	1000
Software Licenses (Open Source)	0
Uncertainty	500
Miscellaneous	500-1000
Total Estimated Cost	4500-5500 BDT

5.4 Complex Engineering Problem

The project Bengali Question Answering system involved addressing various complex engineering problems which is required deep domain knowledge, creative problem solving,

and integration of multidisciplinary skills. We detailed how this project mapped into complex problem solving and engineering activity frameworks in the below.

5.4.1 Complex Problem Solving

The development of Bengali QA system represents a complex engineering problem, containing challenges like fine-tuning large language models, handling Bengali language datasets. Use of pre-trained LLaMA model with LoRA to manage GPU memory constraints. Data training, evaluation using ROUGE metrics helped in addressing the complexities.

Table 5.2: Mapping with complex problem solving

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of ApplicableCodes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓		✓	✓		✓	✓

Justification for EP1: Applied advanced Deep Learning, NLP, and Bangla linguistics to design, train, and evaluate the models.

Justification for EP3: Conducted systematic experimentation with hyperparameters, model architectures, and dataset versions.

Justification for EP4: Addressed multilingual model adaptation, tokenization challenges, and low-resource for Bengali issues.

Justification for EP6: Considered needs of diverse stakeholders including students, researchers, and Bangla-speaking communities.

Justification for EP7: Integrated dataset preparation, fine-tuning, BLEU and ROUGE based evaluation into a single automated workflow using HuggingFace.

Mapping with Knowledge Profile for EP1

This Table 5.3 is designed to map the EP1 to the Knowledge Profile.

Table 5.3: Mapping with knowledge profile

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
		✓	✓	✓	✓		✓

Justifications for each Knowledge Profile element:

K3: Fundamental engineering knowledge in algorithms, data preprocessing, text encoding, model evaluation metrics like BLEU, ROUGE, and software development were applied extensively.

K4: Specialist knowledge in the embeddings, advanced transformer libraries, and the tokenization techniques were applied in the project to overcome this knowledge profile.

K5: Systematic design of throughout architecture from dataset preparation, model fine-tuning, evaluation and a deployed web application was carried out to grain the following engineering design problem.

K6: Engineering practice is obtained with using modern AI tools, deployment, evaluation, documenting the project, limitations and sharing results of the project.

K8: Literature review presented a crucial evaluation of existing models, system recommendation and different technologies used by other researchers. Analyzing and applying published knowledge is also a part of this knowledge profile.

5.4.2 Engineering Activities

In terms of engineering activities, the project mapped across a range of standard categories that define engineering practice complexity.

Table 5.4: Mapping with complex engineering activities

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and	EA5 Familiarity
✓		✓	✓	✓

Justification for attainment EA1: During this project, we utilized wide range of engineering resources including LLM architecture, deep learning frameworks (PyTorch, Transformers), fine-tuning techniques (LoRA, Unsloth), Bengali QA datasets and evaluation metrics (BLEU, ROUGE).

Justification for attainment EA3: We adopted a pre-trained model LLaMA for specific language task like Bengali Question Answering and experimented with different fine-tuning strategies such as LoRA and Unsloth.

Justification for attainment EA4: The project has beneficial social and environmental effects. Providing digital learning resources, it supports educational access and reduces dependency on printed materials.

Justification for attainment EA5: We developed a strong familiarity with modern AI techniques, deep learning method and a large language model architecture.

5.5 Summary

In this chapter, we discussed about how the project agrees with engineering standards, ethical principles and social responsibilities while the system was preparing for real world deployment. Ethical principles like fairness, data privacy and transparency were merged. The deployments potential impact on society was highlighted. Assessment of the system incorporated intricate engineering frameworks to reflect technical depth and practical viability all needed to move to the production environment.

Chapter 6

Conclusion

This chapter gives an overall conclusion to the Bengali QA system project. It provides the summary of the accomplishment reviews the limitations faced during development. It also gives ideas about further improvement to the system. The mail goal of the project is to utilize Natural Language Processing (NLP) methods for Bengali language in the context of AI.

6.1 Summary

This research proposes a methodology for generating answers in Bengali using transformer models named LLaMA. The main objective was to create a scalable system that can generate mostly accurate answers in Bengali from the input question. We initiated studying previous works and started finding gaps in Bangla NLP. To evaluate the findings we have created a structured QA dataset for leveraging contribution in low resource languages.

The research methodology followed a structured process absorbing data collection, preprocessing, model fine-tuning and evaluation. A raw structured dataset of 15000 Bangla QA pairs were prepared to ensure linguistic and context richness. In preprocessing method text normalization, tokenization and dataset division into train test for competent training. Automatic evaluation metrics like BLEU, ROUGE and Exact Match were integrated.

The findings desirable performance by LLaMA with 42.38% Exact Match, 51.36% BLEU scores. The LLaMA is also performed better in ROUGE scores due to language specific pretraining.

The research is meaningfully contribution to Bangla NLP by demonstrating structured approaches and transformers architecture are beneficial to overcome low-resource language processing challenges. In addition, the system preliminary step toward multilingual, domain customizeable and sophisticated Bangla tools with huge potential into educational platform.

6.2 Limitation

In spite of the obtained results, the system has several limitations. First, the dataset contains only 15000 Bengali QA pairs which is relatively so small compared to available English datasets or other rich-resource language. This limited size dataset restricts the models ability.

Second, the model struggles with long or multi sentence questions and inconsistent answers. Third, the fine-tuning process is highly dependent on GPU resources, making experiment with larger models is expensive and time consuming. For domain specific adaption the system limits on the effectiveness in various fields like law, medicine and others.

6.3 Future Work

Based on the existing system success and its limitations, there are diverse directions for future work with increasing its capabilities. The most decent way to increase the dataset. A large diverse domain specific text from literature, history, science can enable model generalization. With accessing more higher computational power, future guidance might include fine-tuning larger models.

Future enhancements will also focus on generating more complex question answers such as descriptive, inference based, and yes or no questions making the system more useful in educational assessment. Additionally, an user friendly mobile application can be deployed. More features like batch input, dataset upload, real-time generation can be implemented.

To ensure the system works on mobile device, model compression techniques such as knowledge distillation can be applied. This method can speed up the execution on limited resources platform like smartphones and tablets. Post processing methods can also be integrated to automatically fix minor grammatical errors for improving the output generation.

At last, realising the system with a liberal license will motivate developers and researchers further development comprehensive adoption. With continuous improvements, the project can be a powerful, scalable and accessible tool.

References

[1] F. Rabbi and M. R. Biswas, "Annotated Bangla Natural Language Processing (BNLP) Using Python and Machine Learning," Maneesha R., Annotated Bangla Natural Language Processing (BNLP) Using Python and Machine Learning (November 28, 2024), 2024.

[2] S. Sengu¹pta, S. Ghosh, P. Mitra, and T. I. Tamiti, "Milestones in Bengali Sentiment Analysis leveraging Transformer-models: Fundamentals, Challenges and Future Directions," arXiv preprint arXiv:2401.07847, 2024.

[3] S. C. Roy and M. M. H. Manik, "Question-Answering System for Bangla: Fine-tuning BERT-Bangla for a Closed Domain," arXiv preprint arXiv:2410.03923, 2024.

[4] H. Mahmud, H. Mahmud, and M. R. A. Rashid, "Enhancing Sentiment Analysis in Bengali Texts: A Hybrid Approach Using Lexicon-Based Algorithm and Pretrained Language Model Bangla-BERT," arXiv preprint arXiv:2411.19584, 2024.

[5] A. Bhattacharjee et al., "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," arXiv preprint arXiv:2101.00204, 2021.

[6] D. Su et al., "Generalizing question answering system with pre-trained language model fine-tuning," in Proceedings of the 2nd workshop on machine reading for question answering, 2019, pp. 203-211.

[7] S. Zhang et al., "Enhanced Fine-Tuning of Lightweight Domain-Specific Q&A Model Based on Large Language Models," in 2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW), 2024: IEEE, pp. 61-66.

[8] M. Engelbach, D. Klau, F. Scheerer, J. Drawehn, and M. Kintz, "Fine-tuning and aligning question answering models for complex information extraction tasks," arXiv preprint arXiv:2309.14805, 2023.

[9] K. Guo, D. Diefenbach, A. Gourru, and C. Gravier, "Fine-tuning Strategies for Domain Specific Question Answering under Low Annotation Budget Constraints," in 2023 IEEE

35th International Conference on Tools with Artificial Intelligence (ICTAI), 2023: IEEE, pp. 166-171.

[10] S. Kim, "Medbiolm: Optimizing medical and biological qa with fine-tuned large language models and retrieval-augmented generation," arXiv preprint arXiv:2502.03004, 2025.

[11] R. Ragab and A. Altahhan, "Fine-tuning of Small/medium LLMs for Business Qa on Structured Data," Available at SSRN 4850031, 2024.

[12] H. Luo et al., "Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models," arXiv preprint arXiv:2310.08975, 2023.

[13] A. Khandelwal, "DomainInv: Domain Invariant Fine Tuning and Adversarial Label Correction For Unsupervised QA Domain Adaptation," in Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), 2024, pp. 13-25.

[14] S. Soni and K. Roberts, "Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering," in Proceedings of the twelfth language resources and evaluation conference, 2020, pp. 5532-5538.

[15] G. Tsatsaronis et al., "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," BMC bioinformatics, vol. 16, no. 1, p. 138, 2015.

213-15-14771

ORIGINALITY REPORT

19%	15%	8%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	6%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	arxiv.org Internet Source	2%
4	Submitted to United International University Student Paper	1%
5	Submitted to Khulna University of Engineering & Technology Student Paper	1%
6	aclanthology.org Internet Source	<1%
7	linnk.ai Internet Source	<1%
8	Submitted to Mawlana Bhashani Science and Technology University Student Paper	<1%
9	www.arxiv.org Internet Source	<1%
10	Submitted to University of York Student Paper	<1%
11	"Artificial Neural Networks and Machine Learning – ICANN 2025", Springer Science and Business Media LLC, 2026 Publication	<1%

12	www.slideshare.net Internet Source	<1%
13	Peter Andrew, Abba Suganda Girsang. "IndoBART optimization for question answer generation system with longformer attention", International Journal of Informatics and Communication Technology (IJ-ICT), 2025 Publication	<1%
14	Submitted to National College of Ireland Student Paper	<1%
15	ia904700.us.archive.org Internet Source	<1%
16	Faisal Ibn Aziz, Muhammad Nazrul Islam. "BanglaHealth: A Bengali paraphrase Dataset on Health Domain", Data in Brief, 2025 Publication	<1%
17	huggingface.co Internet Source	<1%
18	Md Mehrab Hossain, Iffat Ara Helal Akhi, Syeda Rafiatus Sama Raisa, Taskin Sultana Onni, Ariful Islam Rifat, Ashraful Islam. "Critical Analysis of BERT and LSTM Model for Bengali Sentiment Analysis Across Varied Datasets", 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN), 2023 Publication	<1%
19	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using computational intelligence", CRC Press, 2025 Publication	<1%

20	Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, Fumito Masui. "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art", Information Processing & Management, 2023 Publication	<1 %
21	Submitted to University of Hertfordshire Student Paper	<1 %
22	htwk-leipzig.qucosa.de Internet Source	<1 %
23	ijrpr.com Internet Source	<1 %
24	www.coursehero.com Internet Source	<1 %
25	Submitted to University of Mauritius Student Paper	<1 %
26	Submitted to Liverpool John Moores University Student Paper	<1 %
27	perspectivebd.com Internet Source	<1 %
28	Submitted to Berlin School of Business and Innovation Student Paper	<1 %
29	Zhang, Jiyin. "Knowledge-Infused LLM Application in Data Analytics: Using Mindat as an Example", University of Idaho, 2025 Publication	<1 %
30	hdl.handle.net Internet Source	<1 %
31	docs.neu.edu.tr Internet Source	<1 %

32	pythonmania.org Internet Source	<1%
33	"Proceedings of 4th International Conference on Artificial Intelligence and Smart Energy", Springer Science and Business Media LLC, 2024 Publication	<1%
34	Bui Thanh Hung, M. Sekar, Ayhan Esi, R. Senthil Kumar. "Applications of Mathematics in Science and Technology - International Conference on Mathematical Appplications in Science and Technology", CRC Press, 2025 Publication	<1%
35	Md. Jahid Hasan, Md. Ferdous Wahid, Md. Shahin Alom, Mohammad Mahmudul Alam Mia. "A New State of Art Deep Learning Approach for Bangla Handwritten Digit Recognition using SVM Classifier", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020 Publication	<1%
36	s3-eu-west-1.amazonaws.com Internet Source	<1%
37	export.arxiv.org Internet Source	<1%
38	unipress.au Internet Source	<1%
39	www.researchsquare.com Internet Source	<1%
40	"Big Data Analytics in Astronomy, Science, and Engineering", Springer Science and Business Media LLC, 2025	<1%

Publication

41	Jiancheng Du, Yang Gao. "Domain Adaptation and Summary Distillation for Unsupervised Query Focused Summarization", IEEE Transactions on Knowledge and Data Engineering, 2024 Publication	<1%
42	pdfs.semanticscholar.org Internet Source	<1%
43	www.jmir.org Internet Source	<1%
44	Dongmo, Cyrille. "Formalising Non-Functional Requirements Embedded in User Requirements Notation (URN) Models", University of South Africa (South Africa) Publication	<1%
45	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	<1%
46	Beiler, Andrew. "Data-Driven Optimization of Fine-Tuned Large Language Models: Analyzing Data Characteristics for Enhancing Performance With Smaller Base-Models", Saint Louis University, 2024 Publication	<1%
47	Dhruv Kolhatkar, Devika Verma. "Indic Language Question Answering: A Survey", 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2023 Publication	<1%
48	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC	<1%

Press, 2025
Publication

49	Samreen Kazi, Shakeel Khoja, Ali Daud. "A survey of deep learning techniques for machine reading comprehension", Artificial Intelligence Review, 2023 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		