Fine-Tuning Pretrained Language Models for Bengali Poetry Generation

Ankita Porel (2024201043)
Swastik Pal (2024201011)
Venkat Raghav S (2022101013)

April 07, 2025

Abstract

This progress report details the initial stages of our project, "Fine-Tuning Pretrained Language Models for Bengali Poetry Generation." The project aims to enhance the capability of pretrained language models (e.g., Gemma3, GPT-2, Sarvam-AI) to generate culturally resonant and stylistically accurate Bengali poetry. As of Week 1-2, we have defined the project scope, conducted a preliminary literature review, and initiated data collection and preprocessing using the Bengali Poem Dataset from GitHub. This report outlines our objectives, completed tasks, challenges encountered, and next steps.

1 Introduction

Large-scale language models have shown remarkable success in text generation, yet their application to poetry or different languages than English remains limited due to scarce training data, complex poetic forms, and the language's morphological richness. Our project seeks to address these challenges by fine-tuning a pretrained model on a curated Bengali poetry dataset. This report summarizes our progress as of April 07, 2025, aligning with the timeline outlined in the project plan.

2 Objectives

The primary objectives of this project are:

- To fine-tune a pretrained language model (e.g., Gemma-3, GPT-2, Sarvam-AI) for generating coherent and stylistically accurate Bengali poetry.
- To incorporate Bengali poetic structures, such as meter and rhyme, into the generated outputs.
- To evaluate the model's performance using quantitative metrics (BLEU, ROUGE, CHRF) and qualitative human assessments for understanding rhyming sequences.

3 Progress Update

3.1 Problem Understanding & Research (Week 1)

We have completed the following tasks:

- Defined the project scope: Fine-tuning a pretrained model to generate Bengali poetry with cultural and linguistic fidelity.
- Conducted a literature review, studying key works such as:
  - Pascual (2021) on deep learning approaches for poetry generation.
  - Acharya (2024) on fine-tuning Gemma-2 for Bengali poetry.
  - Stanford NLP (2024) on prefix-control in multilingual poetry generation.
- Analyzed Bengali poetic structures, focusing on meter, rhyme, and grammatical norms.

3.2 Data Collection & Preprocessing (Week 1-2)

We have made significant progress in preparing the dataset:

- Acquired the Bengali Poem Dataset from GitHub, containing 6,070 poems by 137 poets.
- Initiated preprocessing:
  - Tokenization: Experimented with SentencePiece for breaking poems into words and subwords suitable for Bengali.
  - Normalization: Converted text to lowercase and removed non-poetic elements (e.g., metadata, annotations).
- Ensured a balanced distribution of classic and modern poems for training and evaluation.

3.3 Model Selection

We are currently evaluating pretrained models:

- Sarvam AI: Considered because it is tailored for many Indian languages like- Bengali, Hindi, Kannda, Malyalam, etc.
- Gemma-3: Considered for potential adaptation to poetry generation in Bengali.

Initial experiments with these models are completed.

4 Challenges Encountered

- **Dataset Limitations**: The Bengali Poem Dataset, while substantial, lacks diversity in certain poetic styles, which may affect model generalization.
- **Tokenizer Selection**: Standard tokenizers struggle with Bengali's morphological complexity, necessitating further experimentation with a few custom tokenizers, such as gemma3, sarvam-ai, etc.
- **Resource Constraints**: Fine-tuning large models like LLaMA requires significant computational resources, which we are currently optimizing.
- **Results**: We did some initial experimentation with the Sarvam AI model. The results of experimentation discussed in more detail in Section 5.

5 Initial Experimentation

5.1 Tokenizers

We experimented with a few custom tokenizers. The results are shown below. All tokenization performed on the same sentence.

- indic-bert Tokenizer

- sarvam-ai's sarvam-1 Tokenizer

- gemma3- 1 billion parameters

- gemma3- 2 billion parameters

Discussion:
It is observed that, tokenization in Sarvam-1 and Gemma3 (1B) gives a word level tokenization, which is more suited for general text-analysis task, and struggles with unknown words and large vocabulary size.
In Gemma3 (2B) and Indic-BERT, the tokens formed gives the impression of subword tokenization. This might be useful, as there are morphological patterns in poems in Bengali, like 'Dhonyatak Shobdo', 'Onukar'.

Check the tokenization results here.

5.2 Sarvam AI (with and without fine-tuning)

We did some initial experimentation with the Sarvam AI model, and evaluated it on certain metrics. We did so with both the base version and the fine-tuned version of the model. The results of those evaluations are given below. (Test is performed on randomly selected 10% of the dataset. This test dataset is kept same for all the tests)
About the Test Metrics:
BLEU (Bilingual Evaluation Understudy): Measures precision of n-gram overlap

between generated and reference text. For a Bengali fine-tuned model, it assesses exact word matches, ranging from 0 to 1 (higher is better).

ChRF (Character n-gram F-score): Evaluates character-level n-gram similarity, balancing precision and recall (0–100). Ideal for Bengali's rich morphology, capturing partial matches beyond word boundaries.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Computes recall-based n-gram overlap (ROUGE-1, ROUGE-2, ROUGE-L), from 0 to 1. For Bengali, it gauges content and sequence similarity, e.g., poetic structure.

These metrics together provide a comprehensive evaluation of the model's Bengali text generation.

- **ROUGE Scores:**
  - ROUGE-1: 0.000016
  - ROUGE-2: 0.000014
  - ROUGE-L: 0.000016
- **BLEU Score:** 0.0171
- **ChRF Score:** 12.4741

5.3 Sarvam AI (with fine-tuning)

- **ROUGE Scores:**
  - ROUGE-1: 0.3210
  - ROUGE-2: 0.1841
  - ROUGE-L: 0.2570
- **BLEU Score:** 0.18510
- **ChRF Score:** 42.6665

Link for the model- Sarvam1 Fine-tuned Model

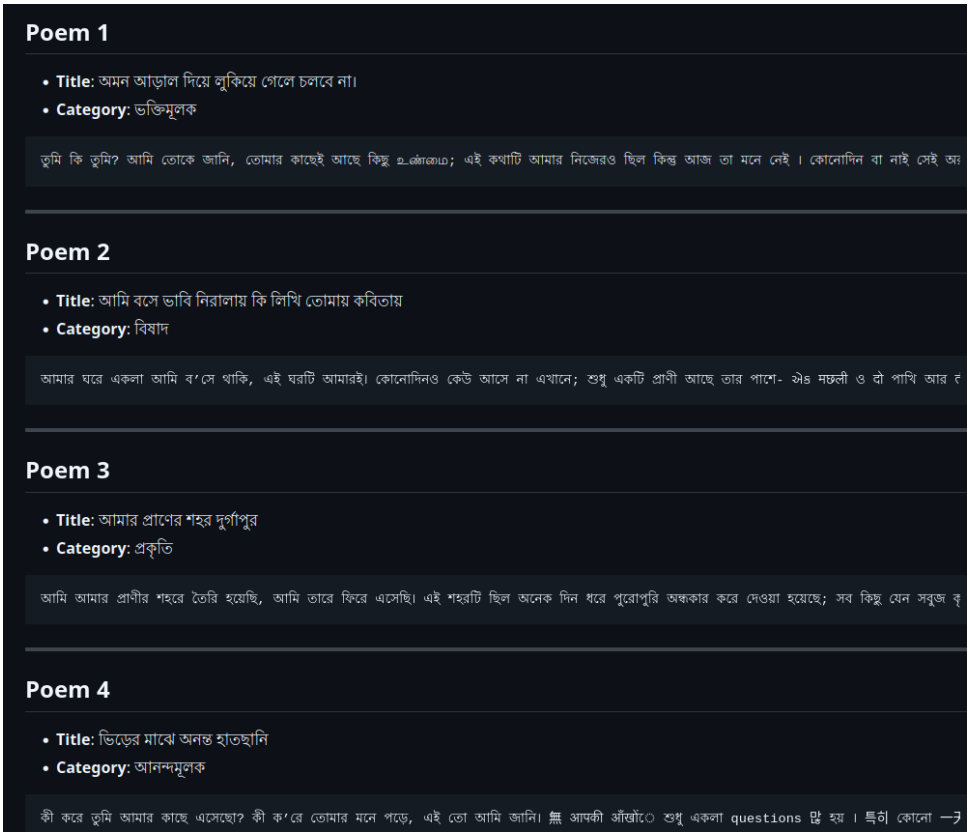
5.4 Interpreting the Results

The ROUGE metric (normalized) ranges from 0 to 1. For ROUGE-1, a score of around 0.5 or above is considered good, and for ROUGE-2 and ROUGE-L, a score of around 0.4 or above is considered good.
The BLEU metric (normalized) ranges from 0 to 1. For this metric score of around 0.5 or above is considered good.
The ChRF metric ranges from 0 to 100. For this metric, a score of around 60 or above is considered good.
The metrics improved a lot after fine-tuning, but they were still below the cutoffs. This means that fine-tuning improved the performance of the model quite a bit, but still not enough to be considered good.

5.5 Gemma-3 1B Fine-Tuning Results

Disclaimer: Due to time constraints on online GPU service platforms, the test metrics for this model is not provided as of now. They will be included in final submission.
The link for our fine-tuned model, that can be used for inference or further training- Gemma3-1b-v1

- Output:

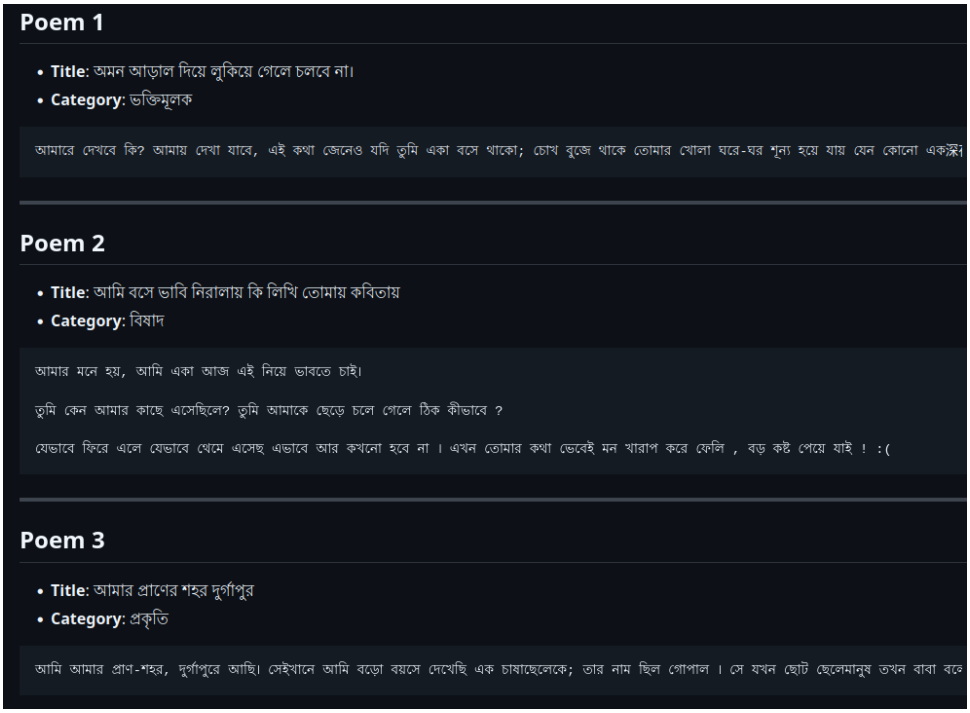


- Github link to check the test runs
  Discussion:
  The poems do not seem to follow the prompt title and category given in instruction very well. The meaning of the output is incoherent to interpret at some places. The output contains some different characters than Bengali. Rhyming ends are not observed.

5.6 Gemma-3 4B Fine-Tuning Results

Disclaimer: Due to time constraints on online GPU service platforms, the test metrics for this model is not provided as of now. They will be included in final submission.
The link for our fine-tuned model, that can be used for inference or further training- Gemma3-4b-v1

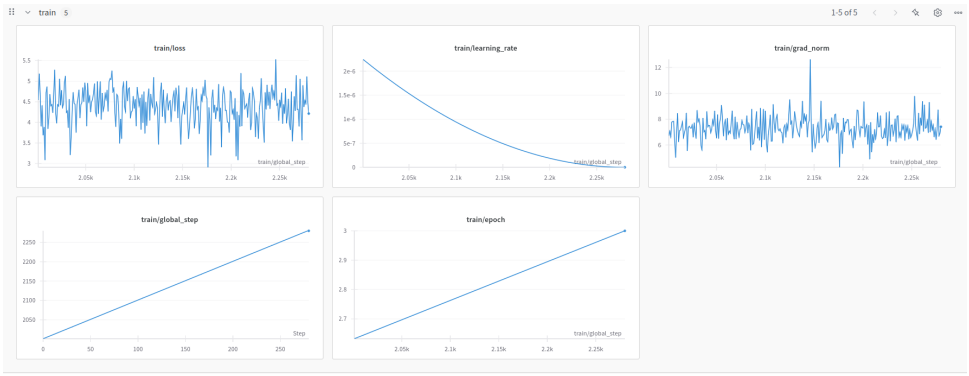
- Output:



- Github link to check the test runs
  Discussion:
  The poems are observed to capture the prompt title and category given in instruction considerably well. Though the output contains a few random characters than Bengali. The poems have interesting patterns, as compared to the one discussed before.

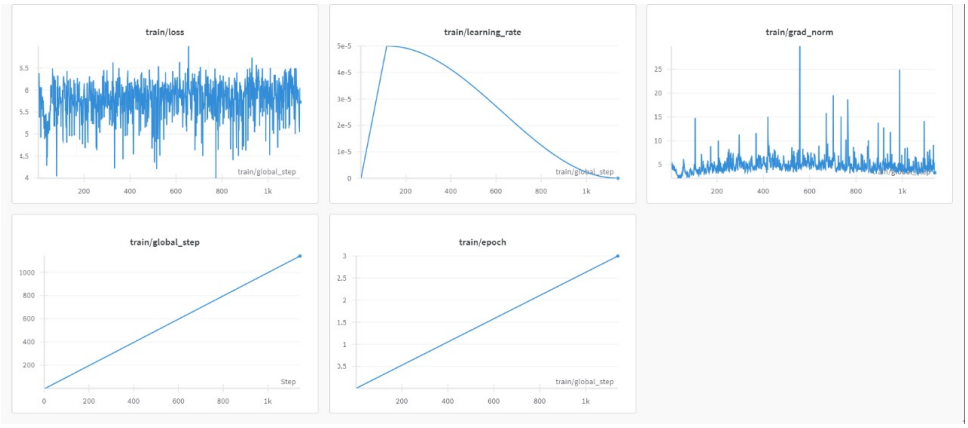
5.7 Visualize weights and biases in training:

- Training loss visualization for Gemma3-4b



- System usage visualization for Gemma3-4b

- Training loss visualization for Gemma3-1b



6 Challenges

- Limitation of GPU compute resources.
- Lack of models that are optimized for Indian Languages.
- Lack of diverse poem dataset in Bengali.
- The existing models do not seem to understand poetic structures very well.

7 Next Steps

Based on the project timeline, our upcoming tasks include:

- Evaluate the metrics for Gemma3 1B and 4B fine-tuned models.
- Conduct initial testing and validate model outputs for fluency and rhythmic structure.