# Can LLMs be Literary Companions?: Analysing LLMs on Bengali Figures of Speech Identification

**Sourav Das**
IIIT Kalyani
sourav_phd21@iiitkalyani.ac.in

**Kripabandhu Ghosh**
IISER Kolkata
kripa.ghosh@gmail.com

## Abstract

Despite Bengali being among the most spoken languages bearing cultural importance and richness, the NLP endeavors on it, remain relatively limited. Figures of speech (FoS) not only contribute to the phonetic and semantic nuances of a language, but they also exhibit aesthetics, expression, and creativity in literature. To our knowledge, in this paper, we present the first ever Bengali figures of speech classification dataset, BengFoS, on works of six renowned poets of Bengali literature. We deploy state-of-the-art (SoTA) models to this dataset, improve them, and finally dissect them, revealing novel insights on the intrinsic behavior of two open-source LLMs (Llama and DeepSeek) in FoS detection. *Though we focused on Bengali, the experimental framework can be reproduced for English as well as for other low-resource languages.* [1]

## 1 Introduction

"হে বঙ্গ ভাণ্ডারে তব বিবিধ রতন; ... মাতৃভাষা-রূপ খনি, পূর্ণ মণিজালে।" (*O Bengali (language) your treasury has various jewels; ... the mother language is a mine full of jewels*).

— *Bongobhasha (The Language of Bengal), 1866, Michael Madhusudan Dutt* (Dutt, 1866)

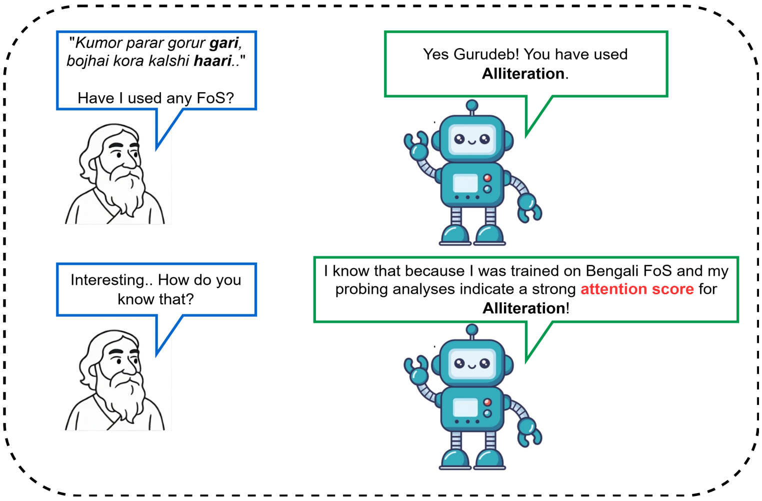
In the above quote, the famous Bengali poet Michael Madhusudan Dutt broods over the richness of the Bengali language in his sonnet *Bongobhasha* (the Bengali language). Notably, the first and last lines of the poem contain figures of speech (FoS), such as apostrophe and Metaphor. Identifying figurative language, particularly FoS, remains a critical yet underexplored challenge. This challenge is even more prominent in low-resource languages, such as Bengali, which ranks as the fifth most widely spoken native language globally (Encyclopaedia Britannica, 2025), yet suffers from a scarcity of annotated linguistic resources and task-specific models. Bengali earned the famous poet Rabindranath Tagore his Nobel prize in 1913 as the first Asian to receive it for his 'Song Offerings' translated from the Bengali poem collection গীতাঞ্জলি (*Gitanjali*). While prior work in Bengali NLP has focused on tasks like sentiment analysis (Sazzed, 2020), named-entity recognition (Ekbal and Bandyopadhyay, 2008), and part-of-speech tagging (Kumar et al., 2024), computational approaches to stylistic determinations like FoS, integral to literary and stylistic analysis, are virtually nonexistent.

FoS detection is vital for applications such as literary critique, machine translation, and creative language understanding, but it presents considerable challenges due to its reliance on subtle semantic and complex contextual cues, particularly in low-resource settings. To address this, we introduce BengFoS, the first systematically curated (to our knowledge), sentence-level dataset for Bengali FoS, comprising 3,148 expertly-annotated literary sentences. Our initial experiments, involving fine-tuning SoTA LLMs such as Llama-3 and DeepSeek R1, underscore the inherent difficulty of Bengali FoS classification, with models achieving only modest gains.

Another distinctive contribution of our research is an in-depth *probing analysis* of these fine-tuned LLMs. While FoS identification has been studied in other languages (e.g., (Berger et al., 2024; Yang et al., 2023)) and LLMs are extensively benchmarked across various NLP tasks like Named Entity Recognition (NER) (Bogdanov et al., 2024) or Question Answering (Li et al., 2024)), such studies predominantly focus on performance metrics. They rarely extend to a detailed investigation of *how* models internalize task-specific knowledge by examining their hidden representations, especially for complex tasks like FoS in less-resourced languages. Our work uniquely bridges this gap.

[1] **Our dataset and codebase are available at:** https://github.com/SouravD-Me/LLMs-on-Bengali-FoS-Identification.

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18634–18656
November 4-9, 2025 ©2025 Association for Computational Linguistics


"Kumor parar gorur gari, bojhai kora kalshi haari.."

Have I used any FoS?
Yes Gurudeb! You have used Alliteration.
Interesting.. How do you know that?
I know that because I was trained on Bengali FoS and my probing analyses indicate a strong attention score for Alliteration!

Figure 1: A light-hearted illustration of an imaginary event where Rabindranath Tagore (image generated by ChatGPT-4o) orates a piece of his poem for children and asks an LLM bot if any FoS is used. The bot responds affirmatively as *Alliteration* (addressing Tagore fondly as *Gurudeb*, his sobriquet), as a classification task, with an explanation by indicating the terminal rhythmic Bengali words in each line, gari (a car) and haari (a pot) through *probing*, which we discuss in detail in the paper.

Following established interpretability methodologies (Jin et al., 2025; Belinkov and Glass, 2019), we apply linear probes to the hidden layers of fine-tuned Llama-3 8B and DeepSeek R1 Distill 7B models. This allows us to assess how figurative language knowledge is encoded and processed within these architectures. This probing, a first for Bengali literary analysis, reveals that FoS-related semantic and lexical cues are distributed across various model depths, with different layers specializing in distinct aspects of figurative language. This deeper understanding of model behavior for FoS identification is a primary focus of our contribution.

Figure 1 imagines a hilarious situation where Rabindranath Tagore interacts with an LLM bot for the FoS in his composed lines. On a more serious note, we look to apply SoTA probing techniques to unearth the reasoning for FoS detection by leading LLMs on our proposed corpus. We make the following contributions through our work:

- We introduce BengFoS, the first gold-standard annotated corpus for Bengali FoS detection, containing 3,148 poetic sentences (Section 2).
- We present a large-scale evaluation of state-of-the-art LLMs (Llama-3 8B and DeepSeek R1 Distill 7B) on the FoS task, including zero-shot baselines, dedicated fine-tuning, and deployment. We report cross-validated performance metrics and detailed comparison results (Section 5).
- We perform in-depth probing analyses of the fine-tuned models, examining their layer-wise representations for FoS knowledge, thereby providing novel insights into how figurative language is internally represented by LLMs (Section 6).
- We analyze the results to identify the challenge that even fine-tuned LLMs yield limited accuracy, and we observe that model attention patterns do not consistently align with human-annotated FoS spans. These findings highlight the need for research focus on linguistically-informed modeling and interpretability of figurative language processing (Section 6).

## 2 Corpus Preparation

We developed the dataset BengFoS by crawling the poems from several internet archives and digital repositories. Only certain poets were selected based on their contrasting writing styles, resource availability, and non-conflicting with the Copyright Act (Copyright Office, Government of India, 1957). Our dataset comprises sentences from poems written by 6 renowned Bengali poets.

Each sentence has been attributed to a specific Sentence ID to keep track of data. We accumulated a total of 10,198 sentences, out of which 3,148 sentences were found to contain FoS and annotated. A reason behind it is that only the poems

| **POET** | **GOLD STANDARD** | **FoS DISTRIBUTION** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| Rabindranath Tagore | 1732 | 384 | 176 | 247 | 278 | 56 | 74 | 813 | 6 | 8 | 5 | 113 | 7 | 1 |
| Jibanananda Das | 345 | 109 | 68 | 79 | 90 | 2 | 12 | 68 | 3 | 0 | 0 | 11 | 0 | 0 |
| Sukanta Bhattacharya | 405 | 133 | 11 | 94 | 81 | 4 | 11 | 184 | 0 | 0 | 0 | 8 | 0 | 0 |
| Kazi Nazrul Islam | 453 | 201 | 15 | 76 | 27 | 3 | 10 | 193 | 1 | 0 | 0 | 12 | 0 | 0 |
| Micheal Madhusudan Dutt | 92 | 58 | 7 | 6 | 8 | 0 | 1 | 50 | 1 | 3 | 4 | 10 | 5 | 0 |
| Sukumar Roy | 121 | 45 | 1 | 10 | 13 | 0 | 2 | 34 | 2 | 5 | 3 | 8 | 4 | 0 |
| **TOTAL** | **3148** | **930** | **278** | **512** | **497** | **65** | **110** | **1342** | **13** | **16** | **12** | **162** | **16** | **1** |
| **PERCENTAGE** | | 29.55% | 8.83% | 16.27% | 15.79% | 2.06% | 3.49% | 42.63% | 0.41% | 0.51% | 0.38% | 5.15% | 0.51% | 0.03% |

Table 1: FoS distribution across poets. The *TOTAL* values indicate the number of sentences where each FoS is present.

contain FoS and not the stories, novels, or articles. Within the 3,148 sentences, 1,732 sentences are from the poems of Rabindranath Tagore, 345 sentences from Jibanananda Das, 405 sentences from Sukanta Bhattacharya, 453 sentences from Kazi Nazrul Islam, 92 sentences from Michael Madhusudan Dutt, and 121 sentences from Sukumar Roy.

The distribution in Table 1 presents the number of sentences containing each specific figures of speech. For instance, out of the 1,732 annotated sentences from Tagore's poem lines, 813 sentences contain instances of Alliteration (FoS Label: 6). The table summarizes the total occurrences of each figures of speech, the overall number of annotated sentences, and the total percentage for each FoS label within the dataset.

Each sentence was manually annotated by two expert linguists for one or more of the following FoS Labels from 0 to 12: None, Simile, Metaphor, Personification, Onomatopoeia, Hyperbole, Alliteration, Oxymoron & Antithesis, Epigram, Irony, Euphemism & Pun, Apostrophe, and Synecdoche & Metonymy. Out of 3,148 sentences in Beng-FoS, approximately 71% contain at least one FoS instance and 29% are labeled None. A small subset (7.2%) contains multiple FoS labels (e.g., both Metaphor and Hyperbole).

### 2.1 Data Splits and Preprocessing

For all experiments, we used the stratified 5-fold cross-validation to ensure robust performance estimates across FoS categories. In each fold, 80% of the data is used for training, 10% for development, and 10% for testing, preserving the overall label distribution. Prior to model input, sentences are normalized using Unicode NFC normalization and tokenized with the native Bengali tokenizer from the Indic NLP (Kakwani et al., 2020) for both the fine-tuned models. We removed leading/trailing whitespace and collapse consecutive spaces. No additional cleaning (e.g., stopword removal) is performed, as FoS often relies on function words.

### 2.2 Human Annotation

We assigned two native Bengali speakers as annotators, who are not authors of this work. They were thoroughly trained with additional examples, as shown in Table 7, under the guidance of senior faculty members of Bengali literature. Upon independent annotation, they achieved high agreement on annotation quality, which was measured using Cohen's $\kappa$ (Ben-David, 2008), resulting in $\kappa = 0.78$, indicating substantial agreement. Table 2 presents three annotation instances with multiple figures of speech, illustrating the complexity of the task. We show all such instances of multiple FoS labels applied to each sentence using a co-occurrence matrix in Figure 8.

## 3 Experimental Setup

We utilized several SoTA models for the FoS identification as a multi-class classification in a zero-shot experimental setup (Hasan et al., 2024). We aimed to comprehensively compare model performances while ensuring computational efficiency. We used the evaluation dataset containing sentences labelled with their corresponding FoS codes. Our evaluation comprises two categories of models: pre-trained models that can be fine-tuned or used directly, and proprietary models accessible through APIs. We select Llama-3 8B (Grattafiori et al., 2024), DeepSeek R1 Distill 7B (Guo et al., 2025), and Mixtral 7B (Jiang et al., 2024) as the local models for evaluation in a zero-shot setup. The API-based models are GPT 3.5 (Ye et al., 2023) and Gemini 1.5 (Team et al., 2024).

### 3.1 Zero-Shot Classification

Given that the models may not have been fine-tuned on our specific classification task, we adopted a zero-shot learning approach. Our objective was to test several open-source and proprietary models on unseen Bengali FoS, to evalu-

| **Sample Sentence** | **Figures of Speech** |
|---|---|
| আমার বুকের তসবির দেখে জল করে টলমল, জল বলে, আমি এরই লাগি কাঁদি গলিয়া হয়েছি জল। (*Seeing the picture on my chest, the water trembles and sways, and says, It is for this that I have wept and turned into water*) | মানবীকরণ (*Personification),* অনুপ্রাস *(Alliteration)* |
| হেমন্তের ঝড়ে আমি ঝরিব যখন পথের পাতার মতো তুমিও তখন আমার বুকের `পরে শুয়ে রবে? (*When, in the autumn storm, I fall like leaves along the path, will you then lie upon my chest?*) | রূপক (*Metaphor),* অতিশয়োক্তি *(Hyperbole)* |
| দুলে ওঠে দিন; শপথমুখর কিষাণ শ্রমিকপাড়া, হাজারে হাজারে মাঠে বন্দরে আজকে দিয়েছে সাড়া। (*The day sways; the oath-chanting farmer and laborer neighborhoods, thousands upon thousands in the fields and ports have responded today*) | মানবীকরণ (*Personification),* অনুপ্রাস *(Alliteration),* অতিশয়োক্তি *(Hyperbole)* |

Table 2: Annotation instances with multiple FoS labels. Complex for LLMs to identify such intricacies.

| **Model** | **Accuracy** | **Avg. Confidence** | **F1 Score** | **Precision** | **Recall** |
|---|---|---|---|---|---|
| Llama-3 8B | 0.4211 | 0.4401 | 0.4178 | 0.4329 | 0.4016 |
| DeepSeek-R1 Distill 7B | 0.4179 | 0.4310 | 0.4023 | 0.4257 | 0.4175 |
| Mixtral 7B | 0.3536 | 0.3817 | 0.3410 | 0.3729 | 0.3386 |
| GPT-3.5 | 0.3647 | N/A | 0.3538 | 0.3790 | 0.3472 |
| Gemini-1.5 | 0.3818 | N/A | 0.3652 | 0.3812 | 0.3590 |

Table 3: Classification performance comparison of different LLMs on zero-shot setup. Confidence scores were not produced by API-based models such as GPT and Gemini. The low values clearly indicate that the pre-trained SoTA LLMs lack the capability to identify Bengali FoS, thus motivating us to fine-tune.

ate whether the models can identify the FoS labels from their pre-training knowledge. We share the zero-shot results in Table 3.

## 4 Experiments

In this section, we present the fine-tuning experiments on DeepSeek R1 and Llama-3, report performance metrics, and compare full fine-tuning with parameter-efficient variants.

### 4.1 Models Fine-Tuning

Before taking the fine-tuning approach, we trained two traditional machine learning algorithms, Support Vector Machines and Multinomial Naive Bayes, on the Gold Standard dataset to observe their performances. The subsequent test and validation results were uninspiring and on par with the zero-shot results. This is further discussed in Appendix A.5. That motivated us to fine-tune both the better-performing open-source models. Based on the performances in the zero-shot setup, we fine-tuned two better-performing models from the competing models, Llama-3 8B and DeepSeek R1 Distill 7B. Llama is proven for its multilingual capabilities, and DeepSeek R1 is the recent SoTA model to exceed the other language models in several benchmarks, including critical thinking. However, the multilingual prowess of DeepSeek R1 is yet to be evaluated in downstream tasks, and hence, we deploy these two models for FoS evaluation.

### 4.2 Fine-Tuning Results

Table 4 summarizes performance on the held-out test folds (5-fold cross-validation). Full fine-tuning of both models yields substantial gains over zero-shot baselines, while adapter and LoRA variants (Whitehouse et al., 2024) achieve competitive performance with far fewer trainable parameters. However, we observed that 8-bit quantization introduces overhead and limitation in achieving competitive multi-label performance in fine-tuning, and 16-bit quantization produces the best overall performance by enhancing 5% to 7% across all the metrics. *Hence, we pursued the rest of the experiments with the 16-bit quantized fine-tuned models*. Hereafter, all the instances of the fine-tuned DeepSeek R1 and Llama-3 versions should be considered as the same.

Full fine-tuning of Llama-3 achieves the best Macro-F1 of 0.14 and Accuracy of 0.35. Among parameter-efficient methods, LoRA performs closest to full fine-tuning, with less than 2% drop in Macro-F1. The fine-tuning loss performances of all Llama-3 and DeepSeek quantized variants are

| Model Variant | Accuracy | Macro-F1 | Micro-F1 |
|---|---|---|---|
| DeepSeek R1 (full) | 0.55 | 0.53 | 0.56 |
| DeepSeek R1 + Adapters | 0.52 | 0.51 | 0.53 |
| DeepSeek R1 + LoRA | 0.51 | 0.50 | 0.52 |
| DeepSeek R1 (16-bit quantized) | **0.55** | **0.54** | **0.56** |
| Llama-3 (full) | 0.55 | 0.53 | 0.55 |
| Llama-3 + Adapters | 0.53 | 0.52 | 0.54 |
| Llama-3 + LoRA | 0.52 | 0.51 | 0.53 |
| Llama-3 (16-bit quantized) | **0.54** | **0.55** | **0.56** |

Table 4: Fine-tuning performance on BengFoS (5-fold CV) by both LLMs. The 16-bit quantized variants achieve marginally superior results.

visualized in Figure 2.

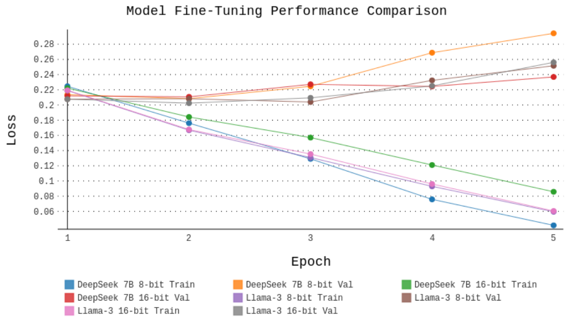


Figure 2: Fine-tuning loss comparison on the gold-standard corpus.

### 4.3 Upsampling of Gold Standard Corpus

To mitigate the label imbalance in our Gold Standard corpus, we also applied a contextual augmentation–based upsampling technique to normalize all FoS labels to the count of the most frequent class. Popular upsampling techniques such as SMOTE (Bhuvana et al., 2025) and ADASYN (Em et al., 2023) operate on numerical feature spaces (generating synthetic feature vectors); they don't directly revert the original natural language text in the corpora or from samples. Applying these techniques typically results in new samples represented as vectors rather than as the original sentences. SMOTE and ADASYN generate synthetic feature vectors that do not directly translate back to the original text. This is even more problematic for a low-resource language like Bengali.

Hence, for our experiment, we utilized the text augmentation technique for negating the label imbalance in the Gold Standard corpus through contextual word embeddings to upsample the minority classes while keeping the original poem sentences in the corpus intact. For this purpose, we used the BERT-based augmented (Şahin, 2022). This augmenter is capable of handling multiple languages (including Bengali) and substitutes words with contextually similar alternatives, preserving original poetic structure while introducing lexical variation. We then *refine-tuned* DeepSeek R1 and Llama-3 in the upsampled corpus and evaluated both finetuning with 8-bit and 16-bit quantized deployment performance. Figure 3 represents a visual understanding of the upsampling of our Gold Standard dataset. The results are discussed in Appendix A.4.

### 4.4 Ablation Studies

To assess the impact of data size and learning rate, we conduct two ablations: **Training Data Fraction:** We fine-tuned both models on 25%, 50%, 75%, and 100% of the training set. **Learning Rate Sweep:** We evaluated learning rates $\{1e{-}5, 3e{-}5, 5e{-}5, 1e{-}4\}$ for full fine-tuning. Ablation analysis demonstrates that performance scales roughly linearly with data size, plateauing beyond 75% of the data, and that $3e{-}5$ remains optimal. These results inform our choice of full data usage and $3e{-}5$ LR for subsequent probing analysis in Section 6.

## 5 Evaluation and Results

We evaluated the fine-tuned system on a held-out set of Bengali sentences with known FoS labels. Additionally, we conduct K-fold cross-validation on the extracted sentence embeddings by training a lightweight logistic regression classifier for multi-label prediction, verifying the consistency of learned representations. We used the standard metrics for multi-label classification evaluation same as the zero-shot setup. The evaluation is shown in Table 5.

### 5.1 Models Deployment

We evaluated the fine-tuned Llama-3 8B and DeepSeek R1 7B models at 16-bit quantized precision on the full BengFoS dataset. The primary objective was to simulate their classification efficacy in a real world deployment scenario. Table 6 presents a comparative summary of their performance metrics.

From the results, the DeepSeek R1 7B (16-bit) model demonstrates notably higher recall and F1-scores across all averaging methods compared to the Llama-3 8B (16-bit) model on this multi-label FoS classification task. For instance, DeepSeek R1 achieved a weighted average F1-score of 0.64, whereas Llama-3 achieved 0.40. While Llama-3

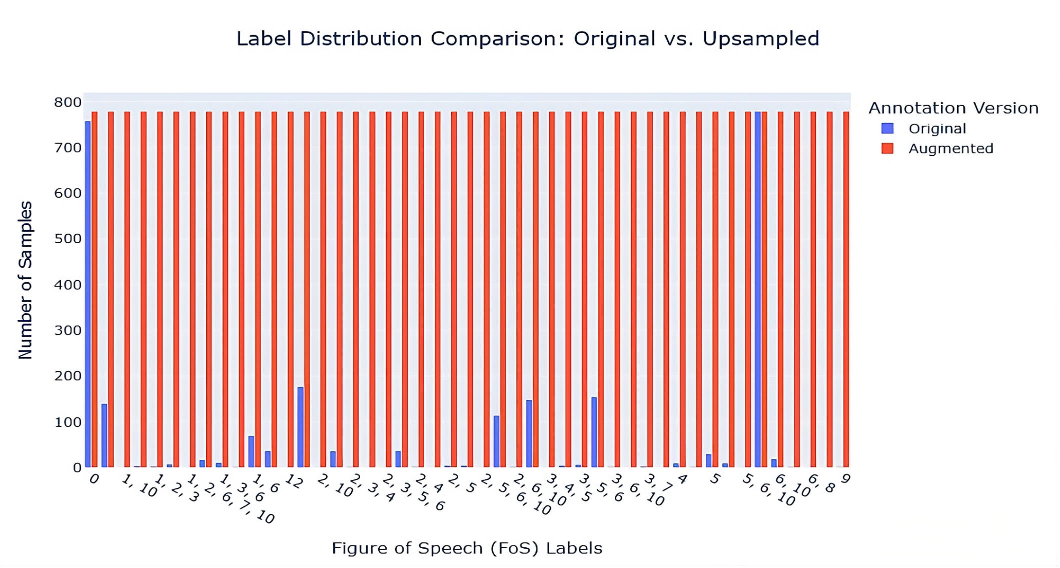


Figure 3: Contextual augmentation upsampling of the Gold Standard corpus. The blue lines represent original annotations bearing corresponding inconsistent FoS labels, while the red lines represent the upsampled FoS labels for all instances.

| **Deployment Metric** | **Llama-3** | **DeepSeek R1** |
|---|---|---|
| **Cross-Validation Averages** | | |
| Accuracy | 0.5500 | 0.5900 |
| Precision | 0.5700 | 0.5800 |
| Recall | 0.6100 | 0.5900 |
| Micro F1 Score | 0.5800 | 0.5900 |
| **Classification Report (Macro Avg)** | | |
| Precision | 0.5700 | 0.6000 |
| Recall | 0.6000 | 0.6500 |
| Macro F1 Score | 0.5700 | 0.6000 |
| **Classification Report (Weighted Avg)** | | |
| Precision | 0.5900 | 0.6200 |
| Recall | 0.6100 | 0.6500 |
| Weighted F1 Score | 0.5900 | 0.6200 |

Table 5: Evaluation of 16-bit quantized models on the full dataset.

shows high recall for certain specific labels (e.g., FoS Labels 4 and 7, based on the detailed classification report), its precision for many classes, and consequently its overall F1-scores, are lower. DeepSeek R1, particularly in terms of recall (e.g., micro avg. recall of 0.92), appears to identify a larger proportion of the true FoS instances, though this sometimes comes at the cost of lower precision for specific minority classes not highlighted in this summary table. The overall performance suggests that for this deployment scenario on the BengFoS dataset, the fine-tuned DeepSeek R1 7B (16-bit) model provides a more effective balance for identifying FoS.

### 5.2 Qualitative Analysis

LoRA-based tuning and quantization on query and value projections align the models to the particularities of Bengali FoS data, resulting in near-parity performance with larger precision variants. In cross-validation experiments, the sentence embeddings yield consistently high micro-averaged precision and recall, confirming that the model captures semantically relevant features of each FoS label. Confusion matrix and ROC curve analyses of both 16-bit quantized models in Appendix A.14 highlight that, quite naturally, the models have performed better in identifying the labels (e.g, Alliteration) they found more than the other labels (e.g, Metonymy).

## 6 Probing Analysis

Language model probing investigates the linguistic knowledge encoded within the internal representations (hidden states) of transformers, typically by training simple linear classifiers on these representations to predict specific properties (Yi et al., 2025; Orgad et al., 2025; Zhao et al., 2024). This technique offers insights into how models process and understand language, moving beyond task performance to interpretability (Conia and Navigli, 2022). In our study, we employ probing to dissect the fine-tuned Llama-3 and DeepSeek R1 Distill

| Model | Llama-3 8B (16-bit) | | | | DeepSeek R1 Distill 7B (16-bit) | | | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **Precision** | **Recall** | **F1-Score** | **Support** | **Precision** | **Recall** | **F1-Score** | **Support** |
| Micro Avg. | 0.17 | 0.53 | 0.26 | 649 | 0.32 | 0.92 | 0.47 | 649 |
| Macro Avg. | 0.15 | 0.49 | 0.19 | 649 | 0.50 | 0.88 | 0.50 | 649 |
| Weighted Avg. | 0.35 | 0.53 | 0.40 | 649 | 0.58 | 0.92 | 0.64 | 649 |
| Samples Avg. | 0.16 | 0.52 | 0.24 | 649 | 0.58 | 0.90 | 0.64 | 649 |

Table 6: Comparative deployment performance of 16-bit quantized models on BengFoS dataset.

models. Specifically, we conduct layer-wise probing using logistic regression to ascertain where and how effectively Bengali FoS categories are represented, analyze hidden state distributions (mean and variance) to understand representational geometry, and examine attention mechanisms. Our goal is to gain a deeper insight of the internal encoding of figurative language within these LLMs for the Bengali FoS identification task.

### 6.1 Setup for Probing

We apply layer-wise linear probing following the methodology (Ju et al., 2024; Cho et al., 2023). After fine-tuning Llama-3 and DeepSeek R1 on the full dataset (using the optimal learning rate and hyperparameters), we freeze the model weights and extract hidden states $\mathbf{h}_\ell$ for each input sentence at every layer $\ell \in \{0, 1, \ldots, L\}$.

We then train a logistic regression probe using 80% of the validation fold to predict the FoS labels based on $\mathbf{h}_\ell$ and evaluate using micro-F1 on the remaining 20%. The probing is performed over all 13 FoS categories. We use the hidden state of the classifier token [CLS] as the sentence-level representation for each layer.

### 6.2 Layer Probing Results

Figure 4 plots the micro-F1 scores achieved by the probing classifier across all 29 layers of DeepSeek R1. We observe that probing performance improves significantly from the embedding layer (Layer 0: 0.292) and peaks around Layer 9 (micro-F1 = 0.575). Beyond this, performance remains relatively stable, with only slight drops in the higher transformer layers. This indicates that FoS-relevant signals are strongly represented in the mid-depth layers of DeepSeek R1.

Interestingly, the middle layers (Layers 8–13) consistently show higher encoding power than the early or final layers. This aligns with previous findings that intermediate layers in transformers capture syntactic and local semantic cues, attributes closely tied to identifying metaphor, personification, or alliteration.

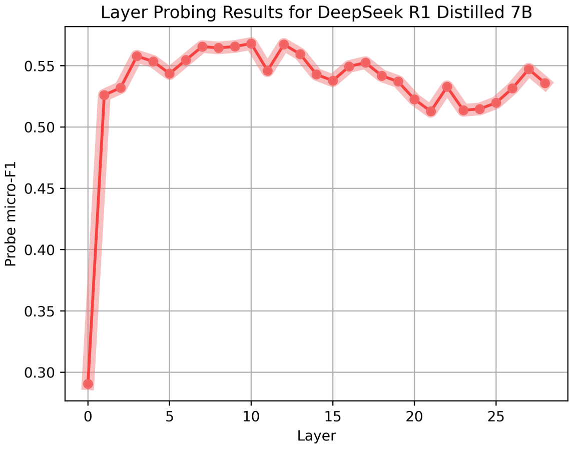


Figure 4: Layer-wise probing results (micro-F1) for DeepSeek R1 fine-tuned on BengFoS.

Figure 5 shows the layer-wise probing trajectory for Llama-3. With 31 layers in total, the model exhibits steady growth in representational quality.

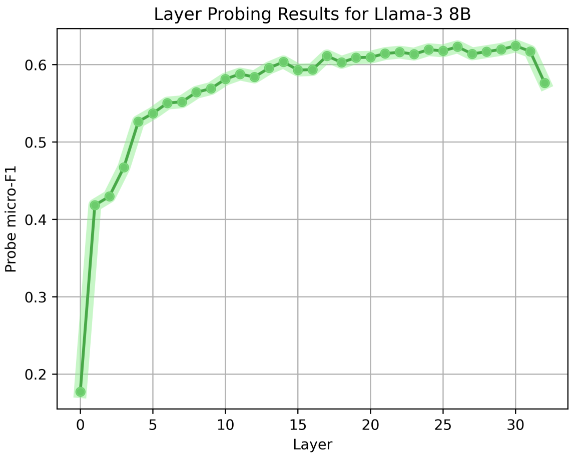


Figure 5: Layer-wise probing results (micro-F1) for Llama-3 fine-tuned on BengFoS.

The probing accuracy begins at a low 0.177 (Layer 0) but increases rapidly, peaking at Layer 30 with micro-F1 = 0.624. Notably, layers 13–30 show a continuous and monotonic rise in performance, suggesting a richer accumulation of abstract semantic features relevant for figurative language. Layer-wise probing analysis revealed distinct peak performance depths for the evaluated models. The DeepSeek R1 model achieved its highest micro-F1 score of 0.575 at Layer 9. In comparison, the Llama-3 8B model demonstrated superior peak performance, reaching a micro-F1 score of 0.624 at a considerably deeper layer, specifically Layer 30.

Our probing analyses share common foundational principles and analogous observations to studies on LLM internal states for statement truthfulness made by Azaria and Mitchell (Azaria and Mitchell, 2023). Both studies affirm that an LLM's internal hidden layer activations are not merely transient computations but encode decipherable information about sophisticated properties, such as statement veracity or nuanced FoS characteristics, suggesting these internal representations hold richer data than surface outputs alone might indicate. Furthermore, a significant common observation is that the decodability of this target information varies across model depth; for instance, specific layers were identified as optimal for truthfulness detection, akin to our findings where FoS classification performance (especially for Llama-3) peaks in the middle to later layers, suggesting a hierarchical processing or refinement of these complex features.

Finally, both analyses hint that an LLM might internally "know" or represent a concept, like a statement being false or the specific features of an FoS, more clearly or robustly in its activations than it might consistently articulate or utilize in its final generated output. These parallels suggest a broader principle: internal state analysis via probing offers a valuable window into the nuanced, and sometimes surprisingly sophisticated, ways LLMs process and represent diverse types of information.

Contrary to the deployment performances, Llama-3 maintains a higher probing peak than DeepSeek R1, indicating better internal abstraction for Bengali FoS classification. DeepSeek's representational strength peaks earlier and plateaus, while Llama-3 continues improving deeper into its stack, possibly benefiting from its larger depth and training corpus. Simultaneously, we also demonstrate a detailed study on the hidden states distributions in Appendix A.10.

### 6.3 Qualitative Insights

We examined a few sentences where both models correctly predicted multiple FoS labels in some and incorrectly in others. One such correctly identified instance is represented in Figure 6 and Figure 7 that shows the token-level attention heatmaps of the same sentence: হাসিয়া উঠিনু ব্যোম পথে, সেথা কেবল শব্দ ওঠে অলখ বাণীর পারাবারে যেন শত শতদল ফোটে। (*Laughing, I rose along the skyward path; there, only a sound stirs on the distant shore of unspoken words, as if hundreds of lotus were bursting into bloom.*) consecutively for DeepSeek R1 and Llama-3 obtained from probing analyses. This sentence contains three FoS labels: Metaphor (Label 2), Hyperbole (Label 5), and Aliteration (Label 6).

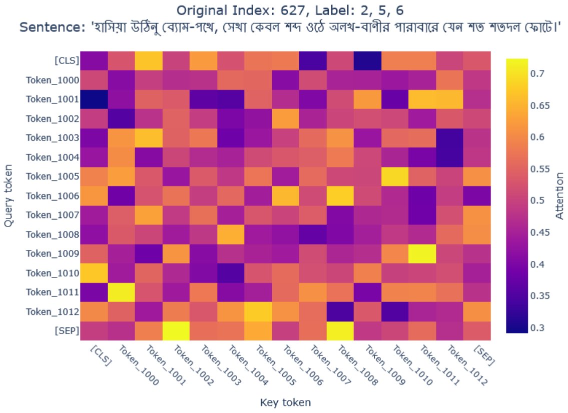


Figure 6: Attention heatmap generated by DeepSeek for a sentence with multiple FoS labels.

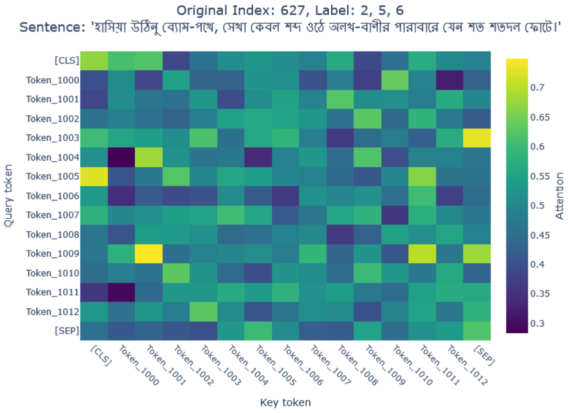


Figure 7: Attention heatmap generated by Llama for a sentence with multiple FoS labels.

Upon inspecting the reason for the correct identification of multi-label FoS within this sentence, we found out that the attention scores in the heatmaps are high for the tokens that fundamentally represent such FoS labels. For instance, considering the first token of the sentence is the classifier token [CLS] and the last token is the separator token [SEP], the tokens responsible for Aliteration are <Token_1006> and [SEP]. For DeepSeek R1's probing analyses, the attention score between them is found to be 0.5045, and for Llama-3's probing analyses, the same is found to be 0.5066. Considering the highest attention score threshold hovered around 0.60 in both the attention heatmaps, respectively, these scores are quite high and indicate that the fine-tuned model found a pattern of rhythmic matching between the words to consider them as the Aliteration. similarly, <Token_1001> and <Token_1002> are responsible for Metaphor, and <Token_1011> and <Token_1012> are responsible for Hyperbole. Both DeepSeek R1 and Llama-3 have captured the linguistic connection between these tokens by capturing the higher attention scores between them than the other remaining tokens. However, there are also several instances where both models failed to identify one or more FoS labels from certain sentences, and we have shown such an example in Appendix A.12.

This analysis indicates that after fine-tuning both models, in several instances, FoS identification by these models is not surface-level pattern matching. It is proven to be deeply entangled with mid-to-high-level semantic composition captured by the transformer layers.

## 7 Conclusion

We introduced BengFoS, the first systematically curated corpus for Bengali FoS, addressing a critical gap for low-resource languages. Extensive evaluations of fine-tuned Llama-3 and DeepSeek R1 highlighted the complexities of FoS identification, with models showing modest overall performance post fine-tuning. Rigorous layer-wise probing revealed how FoS information is encoded across network depths, identifying distinct representational patterns and peak performance layers. Analyses of hidden state distributions and attention mechanisms further clarified this. These novel interpretability studies for Bengali FoS suggest that while LLMs develop internal FoS-indicative representations, translating this to consistently accurate classification remains challenging. Our comprehensive dataset creation, robust model evaluation, and detailed probing advance Bengali literary analysis and also underscore the necessity of integrating interpretability techniques to improve LLMs' capabilities for complex linguistic tasks.

## 8 Limitations

During this work, we faced issues with the unavailability of Bengali literary content on the web as well as copyright-protected content. On one hand, low-resource Indian languages like Bengali have very limited literary resources digitized and restored in any form of digital archives. On the other hand, the Copyright Protection Act of India protects literature and other intellectual properties for a long time, even after the authors' demise. We have discussed this issue in detail in Appendix A.1.

Hence, BengFoS draws from six Bengali poets (authorial skew), and, as a result of their unique writing styles, several FoS classes have low instances. We also initially accumulated the literary works of other famous Bengali authors, such as Sarat Chandra Chattopadhyay, Dwijendralal Ray, and Bibhutibhushan Bandyopadhyay. Despite the fact that most of the digitally available Bengali literary works are not yet outside of copyright, our entire original dataset initially contained 10,198 literary sentences. However, after the accumulation of short stories, novels, and articles written by the above-mentioned authors, the Bengali language experts (who guided us during data annotation) observed that such writings exhibit minimal to no FoS expressions at all. The experts suggested that generally Bengali prose does not contain FoS expressions. That did not align with the objective of this research, and hence, we could not consider those writings for this scope of work.

## 9 Ethical Statement

The annotators were compensated accordingly for their effort on an hourly basis. The proprietary models for the zero-shot learning setup were deployed from Hugging Face without requiring a subscription.

## 10 Acknowledgments

We are grateful to Dr. Shyamasree Biswas Sengupta, Assistant Professor in the Department of Bengali at Kalyani University, India, for her expert guidance in understanding Bengali FoS and the annotation process. We also thank Ms. Sanghamitra Basu and Mr. Subinay Adhikary, both from IISER Kolkata, for their participation in the annotation process.

We are grateful to Mrs. Shyamali Mullick, St. Thomas High School, Dasnagar, Howrah (STHS), India, for her expert guidance in understanding Bengali FoS.

We are also grateful to Dr. Shouvik Kumar Guha, Associate Professor of Law at the West Bengal National University of Juridical Sciences, for guiding us through the Indian Copyright Act for literary works and assisting us in selecting copyright-free literary works for the development of the BengFoS dataset.

Finally, we pay tribute to the late Mr. Santanu Gupta (STHS) for his perennial inspiration and guidance in Bengali literature.

## References

Various Authors. 2025. Bengali literature. https://www.britannica.com/art/Bengali-literature. Last updated May 6, 2025; accessed May 20, 2025.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Arie Ben-David. 2008. Comparison of classification accuracy using cohen's weighted kappa. *Expert Systems with Applications*, 34(2):825–832.

Maria Berger, Sebastian Michael Reimann, and Nieke Marie Kiwitt. 2024. Applying transfer learning to german metaphor prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1383–1392.

J Bhuvana, TT Mirnalinee, Diya Seshan, Avaneesh Koushik, et al. 2025. Ssntrio@ dravidianlangtech 2025: Identification of ai generated content in dravidian languages using transformers. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 335–339.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11829–11841.

Hyunsoo Cho, Hyuhng Joon Kim, Junyeob Kim, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. 2023. Prompt-augmented linear probing: Scaling beyond the limit of few-shot in-context learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12709–12718.

Simone Conia and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632.

Copyright Office, Government of India. 1957. The copyright rules, 1957. Official PDF. Accessed: 2025-09-06.

Michael Madhusudan Dutt. 1866. (bongobhasha). https://www.bangla-kobita.com/madhusudandutt/bongobhasha/. Accessed: 20 May 1866.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Development of bengali named entity tagged corpus and its use in ner systems. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Ranganayaki Em, Abirami Murugappan, Lysa Packiam RS, et al. 2023. Ranganayaki@ lt-edi: Hope speech detection using capsule networks. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 144–148.

Encyclopaedia Britannica. 2025. Bengali language. https://www.britannica.com/topic/Bengali-language. Accessed: 20 May 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, et al. 2025. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 558–573.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8235–8246.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. Part-of-speech tagging for extremely low-resource indian languages. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14422–14431.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18608–18616.

Sharan Narang, Gregory Diamos, Erich Elsen, Paulius Micikevicius, Jonah Alben, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. In *Int. Conf. on Learning Representation*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Gözde Gül Şahin. 2022. To augment or not to augment? a comparative study on text augmentation techniques for low-resource nlp. *Computational Linguistics*, 48(1):5–42.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource bengali language. In *Proceedings of the sixth workshop on noisy user-generated text (W-NUT 2020)*, pages 50–60.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1202–1228.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Xiangpeng Wei, Zhengyuan Liu, and Jun Xie. 2023. Fantastic expressions and where to find them: Chinese simile generation with multiple constraints. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–486.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Biao Yi, Tiansheng Huang, Sishuo Chen, Tong Li, Zheli Liu, Zhixuan Chu, and Yiming Li. 2025. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In *The Thirteenth International Conference on Learning Representations*.

Siyan Zhao, Tung Nguyen, and Aditya Grover. 2024. Probing the decision boundaries of in-context learning in large language models. *Advances in Neural Information Processing Systems*, 37:130408–130432.

## A Appendix

UNESCO observes 21st February as the International Mother Language Day to honor the martyrdom of the Bengalis of the then-Pakistani province of East Bengal (now the independent Bangladesh nation) in the fight for recognition of their Bengali language[4].

The choice of the poets has been made keeping their diverse poetic styles (Authors, 2025). Rabindranath Tagore's writing style stands out for its profound lyricism, philosophical depth, and the seamless blending of Eastern and Western literary traditions. He crafted verse so musical that nearly half his poems became songs, imbuing Bengali literature with a new lyrical idiom. His imagery of nature serves as both setting and symbol for human emotion and universal spirituality.

Jibanananda Das broke entirely new ground in Bengali literature. He invented a fresh poetic diction, fused imagism with existential sensuousness, and pioneered a fragmented, non-linear syntax that demands reading between the lines. His verse is marked by vivid, often surreal imagery of rural Bengal juxtaposed with melancholy and philosophical depth, creating a uniquely modern sensibility rooted in indigenous rhythms and personal memory.

[4] https://en.wikipedia.org/wiki/International_Mother_Language_Day.

Sukanta Bhattacharya's poetry is marked by a revolutionary fervor and proletarian sensibility, blending modernist techniques with accessible, colloquial Bengali to critique social injustice and colonial oppression. His verse employs vivid, concrete imagery drawn from everyday life, hunger, labor, and struggle to evoke both empathy and optimism. Using direct, unadorned diction, he channels Marxist ideals into concise, punchy lines that resonate deeply with the common reader.

Kazi Nazrul Islam revolutionized Bengali literature and music with a style that was simultaneously rebellious, devotional, and deeply humanist. He fused colloquial speech and elevated diction, drawing on Arabic, Persian, Sanskrit, and English registers to craft verse and song that spoke directly to the oppressed masses while engaging classical forms like the ghazal and classical ragas. His language is marked by powerful imagery, rhetorical devices, and innovative metaphors that evoke both the sensual world and lofty ideals of freedom, love, and equality.

Michael Madhusudan Dutt fused Western Romantic and classical influences, especially from Byron, Milton, Homer, and Ovid, with indigenous Bengali rhythms to create an entirely new vernacular poetic idiom. He broke from traditional rhymed Bengali verse by introducing blank verse and was the first to pioneer the sonnet form in Bengali, lending his work both epic grandeur and lyrical intensity.

Sukumar Roy pioneered literary nonsense in Bengali by blending Western influences, notably Edward Lear and Lewis Carroll, with indigenous folk rhythms, creating whimsical yet incisive verse and prose. His language is deceptively simple and colloquial, packed with playful neologisms, puns, and a rhythmic mastery of quatrains and limericks. Beneath the laughter lie subtle social critiques and anti-colonial subtext.

### A.1 Copyright Protection Act

In Section 2 of our paper, we mentioned that only certain poets were selected based on their contrasting writing styles, availability of resources, and adherence to the Copyright Protection Law of India. The Copyright Protection Law of India mandates that a literary work is copyrighted to its author as his/her intellectual property during the author's lifetime and up to 60 years after his/her demise, starting from the beginning of the calendar year immediately following the year of the author's death (Copyright Office, Government of India, 1957). Therefore, we developed the dataset by accumulating the literary works of the six pioneer Bengali poets from the late nineteenth century to the mid-twentieth century, and not beyond that. The dataset curation is governed by the choice of literary works compiled from poem collections, in compliance with the Copyright Protection Law of India.

We carried out our dataset curation in accordance with the Copyright Protection Law of India, as discussed with a senior legal expert *(as mentioned in Section 10)*. We avoided selecting random samples from copyrighted literature and, therefore, curated the data only after discussing potential copyright-related issues that might arise. We have avoided copyrighted materials to safeguard ourselves against potential legal action. Hence, we developed the dataset by accumulating the literary works of the six pioneer Bengali poets from the late nineteenth century to the mid-twentieth century, and not beyond that.

There are 3,148 sentences from poems, and the linguistic experts found FoS labels mostly within these sentences and annotated accordingly. This is our Gold Standard dataset, which is used to fine-tune the Llama-3 and DeepSeek R1 models.

### A.2 Annotation Guidelines

We requested that the annotators independently annotate each sentence with one or more of the 13 possible labels (12 FoS labels and 'None'). They were given a document that contained detailed definitions of FoSs and examples of annotated sentences for each FoS. This document was compiled from a reputed Bangla Grammar book, উচ্চতর বাংলা ব্যাকরণ (*Higher Order Bengali Grammar*) by Bamandev Chakraborty and was finalized by two senior experts in the Bengali language (*as mentioned in Section 10*), who were not otherwise involved in this work. We demonstrate an example of each FoS label annotated for an appropriate literary sentence in Table 7.

### A.3 Zero-shot Classification Discussion

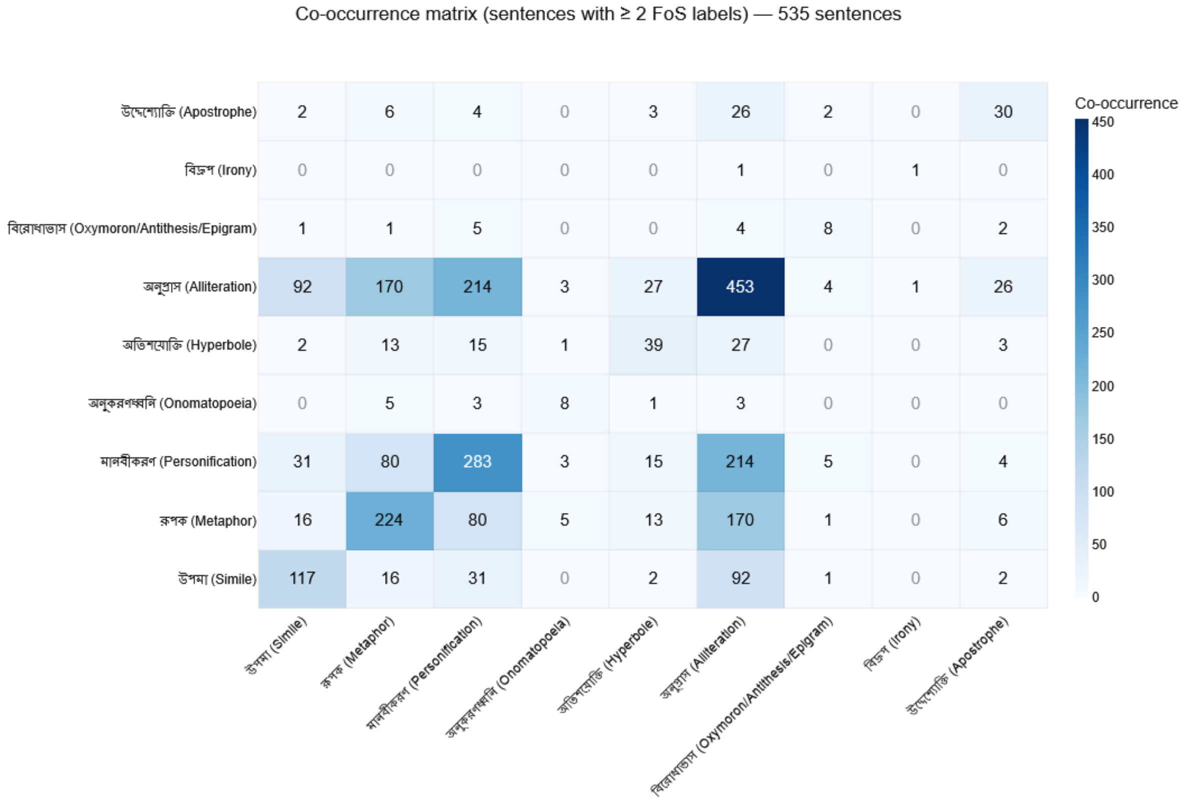
Formally, for an input sentence $x$ and a set of candidate labels $L$, the models compute:

$$\hat{y} = \arg\max_{l \in L} P(l \mid x) \qquad (1)$$

where $P(l|x)$ represents the probability of label $l$ given input $x$ as estimated by the model.

| Class Label | Figures of Speech | Sample Sentence |
|---|---|---|
| 1 | উপমা (*upoma*): An upoma (simile) is a figure of speech that compares an object (*upomeyo*) with another unlike object (*upoman*) using "like", "as", etc. | "ননীর মতো শয্যা কোমল পাতা" (*A bed as soft as butter*) — Kalidas Ray |
| 2 | রূপক (*rupak*): A rupak (similar to metaphor) is a figure of speech that compares an object (*upomeyo*) with another unlike object (*upoman*) such that the latter overshadows the former to the extent that the verb follows the latter. | "জীবন-উদ্যানে তোর যৌবনকুসুমভাতি কতদিন রবে?" (*How long will the bloom of your youth remain in the garden of life?*) — Michael Madhusudan Dutt |
| 3 | মানবীকরণ *manabikaran*: A manabikaran (Personification) is imparting human qualities or abilities to animals or objects. | "যেদিন পূর্ণিমা রাতি আসে চাঁদ আকাশ জুড়িয়া হাসে।" (*The day full moon night comes, the moon smiles across the sky*) — Rabindranath Tagore |
| 4 | অনুকরণধ্বনি (anukarandhwani): An anukarandhwani (Onomatopoeia) uses words that imitate the sounds associated with the objects or actions they refer to. | "মুক্ত যাহার বাণী শুনি কাঁদে ত্রিভুবন থরথর!" (*Hearing whose free-spoken words, the three worlds tremble and weep*) — Kazi Nazrul Islam |
| 5 | অতিশয়োক্তি (atishoyokti): An atishoyokti (Similar to hyperbole) is an exaggeration used for a difference or distinction. | "হাজার হাজার শহীদ ও বীর স্বপ্নে নিবিড় স্মরণে গভীর ভুলি নি তাদের আত্মবিসর্জন।" (*Thousands and thousands of martyrs and heroes, in dreams and deep remembrance, I cherish, I have not forgotten their self-sacrifice*) — Sukanta Bhattacharya |
| 6 | অনুপ্রাস (anupras): An anupras (Alliteration) is the repetition of initial consonant sounds in a series of words. | "আমার বাঁশির সুরে সাড়া তার জাগিবে তখনি" (*The response to the melody of my flute will awaken in them instantly*) — Rabindranath Tagore |
| 7 | বিরোধাভাস (birodhavash): A birodhavash (Similar to oxymoron and antithesis) juxtaposes contrasting ideas in balanced phrases. | "জীবন মৃত্যু পায়ের ভৃত্য, চিত্ত ভাবনাহীন" (*Life and death are servants at my feet, the mind is free from worry*) — Rabindranath Tagore |
| 8 | বিদ্রূপ (bidrup): A bidrup (Irony) conveys a meaning opposite to the literal meaning in a satirical manner. | "কাদম্বিনী মরিয়া প্রমাণ করিল, সে মরে নাই" (*By dying Kadambini proved that she had not died*) — Rabindranath Tagore |
| 9 | শ্লেষ এবং যমক (slesh and yamak): A slesh and yamak (Similar to euphemism and pun) use mild or indirect words with respective positions and multiple meanings to replace harsh or blunt ones. | "কোন গুণ নাই তার কপালে আগুন" (*(He) has no qualities, has fire on forehead*) — Bharatchandra Rai Gunakar |
| 10 | উদ্দেশ্যোক্তি (uddeshyokti): An uddeshyokti (Apostrophe) addresses an absent or imaginary person or a personified abstraction. | "কলকাতা একদিন কল্লোলিনী তিলোত্তমা হবে; তবুও তোমার কাছে আমার হৃদয়।" (*Kolkata you will one day become a bustling, magnificent city; yet, my heart will always belong to you*) — Jibanananda Das |
| 11 | প্রতিনিধিত্ব (pratinidhitwa): A pratinidhitwa (Synecdoche) is a figure of speech in which a part of something is used to represent the whole, or vice versa. | "ক্ষুধার রাজ্যে পৃথিবী গদ্যময়" (*The earth is prosaic in the world of hunger*) — Sukanta Bhattacharya |
| 12 | স্বরবৈশিষ্ট্য (swarbaishistya): A swarbaishistya (Assonance) is a figure of speech in which similar vowel sounds are repeated in nearby words. | "আমি চিরদুর্দম, চিরদুর্মুখ, চিরকুটিল পথ চির!" (*I am ever indomitable, ever defiant, forever treading the crooked path*) — Kazi Nazrul Islam |

Table 7: Examples of all figures of speech labels in Bengali literature.



| | উপমা (Simile) | রূপক (Metaphor) | মানবীকরণ (Personification) | অনুকরণধ্বনি (Onomatopoeia) | অতিশয়োক্তি (Hyperbole) | অনুপ্রাস (Alliteration) | বিরোধাভাস (Oxymoron/Antithesis/Epigram) | বিদ্রুপ (Irony) | উদ্দেশ্যোক্তি (Apostrophe) |
|---|---|---|---|---|---|---|---|---|---|
| উদ্দেশ্যোক্তি (Apostrophe) | 2 | 6 | 4 | 0 | 3 | 26 | 2 | 0 | 30 |
| বিদ্রুপ (Irony) | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| বিরোধাভাস (Oxymoron/Antithesis/Epigram) | 1 | 1 | 5 | 0 | 0 | 4 | 8 | 0 | 2 |
| অনুপ্রাস (Alliteration) | 92 | 170 | 214 | 3 | 27 | 453 | 4 | 1 | 26 |
| অতিশয়োক্তি (Hyperbole) | 2 | 13 | 15 | 1 | 39 | 27 | 0 | 0 | 3 |
| অনুকরণধ্বনি (Onomatopoeia) | 0 | 5 | 3 | 8 | 1 | 3 | 0 | 0 | 0 |
| মানবীকরণ (Personification) | 31 | 80 | 283 | 3 | 15 | 214 | 5 | 0 | 4 |
| রূপক (Metaphor) | 16 | 224 | 80 | 5 | 13 | 170 | 1 | 0 | 6 |
| উপমা (Simile) | 117 | 16 | 31 | 0 | 2 | 92 | 1 | 0 | 2 |

Figure 8: Co-occurrence matrix of FoS labels that are simultaneously applied to the same sentences.

For local models, each model is loaded with the specified quantization configurations to optimize resource usage. Then the input sentences are tokenized and converted into embeddings. The model computes similarity scores between the input embeddings and label representations, estimating $P(l|x)$ for each label $l \in L$. The label with the highest probability is selected as the predicted label. The associated probability $P(y^{|}x)$ serves as the confidence score for the prediction. For API-based models, we interacted with the models by transferring the sentences from the evaluation set to the models. Since API models may not provide explicit confidence scores, we retrieved the other standard set of metrics based on available information.

After evaluating each model, we obtain the confidence scores and compute the performance metrics. We observe that models like Llama-3 and DeepSeek R1 performed competitively. Llama-3 marginally outperformed other models, attributed to its extensive multilingual pre-training data and language understanding capabilities.

While zero-shot classification is powerful, it may not achieve the same performance as models fine-tuned on specific datasets. Here, the models relied on their pre-trained knowledge and hence could not cover domain-specific nuances present in the dataset.

### A.4 Upsampling Experiment Continuation

As shown in Table 8, upsampling yields micro-F1 = 0.4500 and weighted-F1 = 0.4250 for refine-tuned DeepSeek R1, and micro-F1 = 0.5000 and weighted-F1 = 0.5000 for refine-tuned Llama-3. Macro-F1 is 0.4273 for DeepSeek R1 and 0.5000 for Llama-3. On the other side, quantized fine-tuning and deployment results on the original Gold Standard dataset represent better performance, indicating that the upsampling gains are either not preserved or not learned by the internal representations of the language models. Hence, we did not make this a fundamental part of our main experimental framework.

While contextual upsampling can enhance overall classification recall, it does not significantly impact the models' ability to discriminate intricate FoS categories. These findings suggest that further work, such as incorporating synthetic discourse contexts or span-level augmentation, may be neces-

| Model & Mode | Micro-F1 | Macro-F1 | Weighted-F1 | Support |
|---|---|---|---|---|
| DeepSeek R1 7B (fine-tune) | 0.4500 | 0.4273 | 0.4250 | 2,597 |
| DeepSeek R1 7B (8-bit deploy) | 0.4000 | 0.4000 | 0.4000 | 2,597 |
| DeepSeek R1 7B (16-bit deploy) | 0.4250 | 0.4091 | 0.4250 | 2,597 |
| Llama-3 8B (fine-tune) | 0.5000 | 0.5000 | 0.5000 | 2,597 |
| Llama-3 8B (8-bit deploy) | 0.4750 | 0.5000 | 0.5000 | 2,597 |
| Llama-3 8B (16-bit deploy) | 0.4750 | 0.5000 | 0.5000 | 2,597 |

Table 8: Fine-tuning and quantized deployment results after upsampling. The results are subpar to the fine-tuning and deployment on the original Gold Standard dataset, which motivated us to work with that dataset.

sary to improve performance on underrepresented labels.

### A.5 Traditional ML Classifiers Approach

As discussed in Section 4.1, standard SVM and NB classifiers, with bag of words and TF-IDF features, respectively, achieved weighted-F1 scores that are indeed comparable to our zero-shot evaluation setup. However, they fell substantially short of the moderate performance gains by finetuned Llama3 and DeepSeek R1. We show these results in Table 10.

| Model | Micro-F1 (%) |
|---|---|
| SVM (Bag-of-Words) | 41.1 |
| Naïve Bayes (TF-IDF) | 40.5 |
| Zero-Shot Llama-3 | 41.0 |
| Zero-Shot DeepSeek R1 | 40.0 |
| Fine-Tuned Llama-3 | 59.0 |
| Fine-Tuned DeepSeek R1 | 62.0 |

Table 9: Comparison of approaches on the weighted-F1 metric.

### A.6 Data and Task Formulation for Models Fine-Tuning

Our primary goal is to handle FoS assignments, where each sentence may belong to one or more labels. We adopt Low-Rank Adaptation (LoRA) to deploy the models with reduced computational overhead efficiently. For fine-tuning the Llama-3 8B and DeepSeek R1 Distill 7B models on a multi-label classification task involving Bengali text, We consider the GS where each Bengali sentence is associated with one or multiple FoS labels, thus framing the problem as a multi-label classification challenge.

Let:

$$\mathcal{D} = \left\{ \left( x^{(n)}, \mathbf{y}^{(n)} \right) \right\}_{n=1}^{N} \qquad (2)$$

Determining our Gold Standard Bengali dataset of size $N$. Here, $x^{(n)}$ is the n-th sentence, and $y(n) \in \{0,1\}K$ is the corresponding multi-label vector indicating membership in any of $K$ possible FoS categories. We partition $\mathcal{D}$ into training and evaluation subsets, employing a stratified split to preserve label distributions.

Although both models are primarily designed for language modeling tasks, we attach a multi-label classification head to the output of its final attention layer. Concretely, for a sentence $x^{(n)}$, the model outputs a final hidden representation $h^{(n)}$, which we pass through a linear mapping:

$$\mathbf{z}^{(n)} = \mathbf{W}_{\text{cls}}, \mathbf{h}^{(n)} + \mathbf{b}_{\text{cls}}, \qquad (3)$$

where $z^{(n)} \in RK$ is a logit vector for the $K$ FoS labels. Each logit $zk^{(n)}$ is then passed through a sigmoid to yield an independent label probability:

$$p(y_k^{(n)} = 1 \mid x^{(n)}) = \sigma, (z_k^{(n)}). \qquad (4)$$

This setup naturally supports multi-label decisions, allowing each label dimension to be activated independently.

We define our training objective using the Binary Cross-Entropy (BCE) loss over each label:

$$\ell_{\text{BCE}}^{(n,k)} = -y_k^{(n)} \log(\sigma(z_k^{(n)})) - (1 - y_k^{(n)}) \log(1 - \sigma(z_k^{(n)})). \qquad (5)$$

The overall loss is the average across all $n$ and $k$:

$$\mathcal{L} = \frac{1}{N, K}, \sum_{k=1}^{K} \ell_{\text{BCE}}^{(n,k)}. \qquad (6)$$

We minimize $\mathcal{L}$ with an 8-bit variant of the Adam optimizer, following a mixed-precision training scheme (Narang et al., 2017). This setup is particularly suitable when deploying large models in environments with limited GPU memory.

After training, the adapted models can be used to classify new Bengali sentences by computing the final logits for each label and thresholding the sigmoid outputs. Additionally, we extract sentence embeddings by taking the final hidden state or a designated token representation from the last transformer layer:

$$\mathbf{e}^{(n)} = \mathrm{Enc}(x^{(n)}), \tag{7}$$

where *Enc* denotes the final encoder outputs of Llama-3. These embeddings can serve for ancillary tasks such as similarity search or external classifier training.

### A.7 Hyperparameter Search Space

We performed grid search over learning rates $\{1e{-}5, 3e{-}5, 5e{-}5\}$, batch sizes $\{16, 32\}$, and up to 10 epochs with early stopping (patience = 2) using AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay = 0.01). Sequence length was capped at 128 tokens. For adapters, we used a bottleneck d = 64, for LoRA rank r = 8 and r=16, and for prefix-tuning a prompt length of 50 tokens. The insights are represented in Table 10).

| Hyperparameter | Values |
|---|---|
| Learning rate | 1e-5, 3e-5, 5e-5 |
| Batch size | 16, 32 |
| Max epochs | 10 (early stop patience=2) |
| Weight decay | 0.01 |
| Optimizer | AdamW ($\beta_1$=0.9, $\beta_2$=0.999) |
| Sequence length | 256 |
| Adapter bottleneck dim. ($d$) | 64 |
| LoRA rank ($r$) | 8 |
| Prefix length | 50 |

Table 10: Hyperparameter search grid.

### A.8 Fine-Tuning and Deployment Results

The tables below report the entire fine-tuning and deployment metrics (8-bit and 16-bit quantized) for Llama-3 8B and DeepSeek R1 Distill 7B as observed during the experiments. The fine-tuning figures (Accuracy / Macro-F1 / Micro-F1) are taken from Table 4 (5-fold CV fine-tuning variants). The deployment / average metrics (accuracy/precision/recall/micro/macro/weighted F1) are taken from the summary of the deployment evaluation and classification report of the article (Tables 6-7 and related text).

### A.9 Layer-wise Probing Procedure

After fine-tuning, we extracted for each sentence $x_i$ the hidden vector $h_i^{\ell}$ corresponding to the first token (e.g., <s>) at each layer $\ell$. We then trained a separate logistic regression probe in 80% of the validation fold and evaluated micro-F1 on the remaining 20%. Algorithm 1 details the pipeline.

**Algorithm 1** Layer-wise Probing for FoS Label Identification.

```
Require: Fine-tuned transformer model $\mathscr{M}$ with $L$ layers, dataset $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$
for $\ell = 0, \ldots, L$ do
    Extract representations:
        $H^{(\ell)} \leftarrow \{ h_i^{(\ell)} = \mathscr{M}_{\ell}(x_i)[< \mathrm{s} >] \mid (x_i, y_i) \in \mathscr{D}\}$
    Split $(H^{(\ell)}, Y)$ into train (80%) and val (20%)
    Train logistic regression probe $\mathscr{P}^{(\ell)}$ on train
    Evaluate:
        $\mathrm{microF1}^{(\ell)} \leftarrow \mathscr{P}^{(\ell)}(\mathrm{val})$
end for
return $\{\mathrm{microF1}^{(\ell)}\}_{\ell=0}^{L}$
```

### A.10 Analysis of Hidden State Distributions

To further investigate the internal representational dynamics and complement our layer-wise probing accuracy, we analyzed the mean and variance of the hidden state activations for each layer. For a given layer $l$, let $H^{(l)} = \{h_1^{(l)}, h_2^{(l)}, \ldots, h_N^{(l)}\}$ be the set of $N$ hidden state vectors (typically the output of the attention mechanism or the feed-forward network, before layer normalization, averaged over token positions per sentence). We compute the mean vector $\mu^{(l)} = \frac{1}{N}\sum_{i=1}^{N} h_i^{(l)}$ and the mean of the element-wise variances $\sigma^{2,(l)} = \mathrm{mean}(\mathrm{Var}(H^{(l)}))$, where the variance is computed across the dataset for each dimension of the hidden state and then averaged. This analysis aims to reveal layer-specific shifts in representational geometry. For instance, significant changes in variance might indicate layers where representations become more specialized or discriminative for the downstream task.

**DeepSeek R1 Distill 7B Insights:** The DeepSeek R1 model shows in Figure 9 relatively stable variance across most layers, with only a slight increase observed in the final few layers. The mean of its hidden states presents more fluctuation, particularly a sharp decrease in the terminal layers. Unlike Llama-3, a direct and strong correlation between these distributional statistics and the layer-wise probing performance is less apparent for DeepSeek R1, where probing accuracy is more varied across its depth. This difference suggests distinct representational learning strategies between the two architectures when fine-tuned for Bengali FoS identification.

| ID | Figure of Speech (FoS) |
|---|---|
| 0 | None |
| 1 | উপমা (*Simile)* |
| 2 | রূপক (*Metaphor)* |
| 3 | মানবীকরণ (*Personification)* |
| 4 | অনুকরণধ্বনি (*Onomatopoeia)* |
| 5 | অতিশয়োক্তি (*Hyperbole)* |
| 6 | অনুপ্রাস (*Alliteration)* |
| 7 | বিরোধাভাস (*Oxymoron, Antithesis, or Epigram)* |
| 8 | বিদ্রূপ (*Irony)* |
| 9 | শ্লেষ / যমক (*Euphemism or Pun)* |
| 10 | উদ্দেশ্যোক্তি (*Apostrophe)* |
| 11 | প্রতিনিধিত্ব (*Synecdoche and Metonymy)* |
| 12 | স্বরবৈশিষ্ট্য (*Assonance)* |

Table 11: FoS ID → label-name mapping used in the annotations and experiments.

| | **L-3 8-bit** | | | | **L-3 16-bit** | | | | **DS R1 8-bit** | | | | **DS R1 16-bit** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Class** | **P** | **R** | **F1** | **Sup** | **P** | **R** | **F1** | **Sup** | **P** | **R** | **F1** | **Sup** | **P** | **R** | **F1** | **Sup** |
| 0 | 0.66 | 0.60 | 0.63 | 147 | 0.67 | 0.68 | 0.67 | 147 | 0.74 | 0.50 | 0.60 | 147 | 0.74 | 0.50 | 0.60 | 147 |
| 1 | 0.80 | 0.07 | 0.13 | 56 | 0.60 | 0.11 | 0.18 | 56 | 0.50 | 0.09 | 0.15 | 56 | 0.50 | 0.09 | 0.15 | 56 |
| 2 | 0.33 | 0.05 | 0.09 | 81 | 0.42 | 0.14 | 0.21 | 81 | 0.30 | 0.09 | 0.13 | 81 | 0.30 | 0.09 | 0.13 | 81 |
| 3 | 0.30 | 0.06 | 0.10 | 98 | 0.47 | 0.08 | 0.14 | 98 | 0.38 | 0.35 | 0.36 | 98 | 0.38 | 0.35 | 0.36 | 98 |
| 4 | 0.00 | 0.00 | 0.00 | 3 | 0.00 | 0.00 | 0.00 | 3 | 0.00 | 0.00 | 0.00 | 3 | 0.00 | 0.00 | 0.00 | 3 |
| 5 | 0.00 | 0.00 | 0.00 | 15 | 0.00 | 0.00 | 0.00 | 15 | 0.00 | 0.00 | 0.00 | 15 | 0.00 | 0.00 | 0.00 | 15 |
| 6 | 0.69 | 0.73 | 0.71 | 239 | 0.80 | 0.77 | 0.78 | 239 | 0.69 | 0.75 | 0.72 | 239 | 0.69 | 0.75 | 0.72 | 239 |
| 7 | 0.00 | 0.00 | 0.00 | 5 | 0.00 | 0.00 | 0.00 | 5 | 0.00 | 0.00 | 0.00 | 5 | 0.00 | 0.00 | 0.00 | 5 |
| 8 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 |
| 9 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 | 1 | 0.00 | 0.00 | 0.00 | 1 |
| 10 | 0.00 | 0.00 | 0.00 | 21 | 1.00 | 0.05 | 0.09 | 21 | 0.00 | 0.00 | 0.00 | 21 | 0.00 | 0.00 | 0.00 | 21 |
| 11 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 |
| 12 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 |
| **mic. avg** | 0.65 | 0.41 | 0.51 | 666 | 0.69 | 0.46 | 0.54 | 666 | 0.62 | 0.45 | 0.52 | 666 | 0.62 | 0.45 | 0.55 | 666 |
| **mac. avg** | 0.23 | 0.13 | 0.14 | 666 | 0.33 | 0.15 | 0.15 | 666 | 0.22 | 0.15 | 0.16 | 666 | 0.22 | 0.15 | 0.17 | 666 |
| **w. avg** | 0.54 | 0.41 | 0.43 | 666 | 0.64 | 0.46 | 0.46 | 666 | 0.55 | 0.45 | 0.47 | 666 | 0.55 | 0.45 | 0.50 | 666 |
| **smp. avg** | 0.51 | 0.45 | 0.46 | 666 | 0.56 | 0.50 | 0.49 | 666 | 0.50 | 0.47 | 0.47 | 666 | 0.50 | 0.47 | 0.50 | 666 |

Table 12: Fine-tuning — full classification reports (classes 0–12) for the four quantized variants. L-3 is Llama-3 8B, and DS R1 is DeepSeek R1 7B. All metrics are represented from the training logs. Precision, Recall, F1-Score, and Support are represented as P, R, F1, and Sup. The cumulative results are represented in micro average, macro average, weighted average, and sample average.

| | **L-3 8-bit** | | | | **L-3 16-bit** | | | | **DS R1 8-bit** | | | | **DS R1 16-bit** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Class** | **P** | **R** | **F1** | **Sup** | **P** | **R** | **F1** | **Sup** | **P** | **R** | **F1** | **Sup** | **P** | **R** | **F1** | **Sup** |
| 0 | 0.37 | 0.47 | 0.42 | 152 | 0.36 | 0.49 | 0.44 | 152 | 0.96 | 0.66 | 0.79 | 152 | 0.98 | 0.66 | 0.84 | 152 |
| 1 | 0.00 | 0.00 | 0.00 | 51 | 0.00 | 0.00 | 0.01 | 51 | 0.18 | 1.00 | 0.30 | 51 | 0.18 | 1.00 | 0.31 | 51 |
| 2 | 0.19 | 0.78 | 0.31 | 80 | 0.19 | 0.78 | 0.33 | 80 | 0.28 | 1.00 | 0.43 | 80 | 0.28 | 1.00 | 0.46 | 80 |
| 3 | 0.29 | 0.51 | 0.37 | 86 | 0.29 | 0.51 | 0.39 | 86 | 0.29 | 1.00 | 0.45 | 86 | 0.29 | 1.00 | 0.48 | 86 |
| 4 | 0.00 | 0.00 | 0.00 | 4 | 0.01 | 1.00 | 0.02 | 4 | 1.00 | 0.75 | 0.86 | 4 | 1.00 | 0.75 | 0.92 | 4 |
| 5 | 0.03 | 1.00 | 0.05 | 14 | 0.00 | 0.00 | 0.01 | 14 | 1.00 | 0.50 | 0.67 | 14 | 0.05 | 1.00 | 0.11 | 14 |
| 6 | 0.52 | 0.64 | 0.58 | 247 | 0.52 | 0.64 | 0.62 | 247 | 0.67 | 1.00 | 0.80 | 247 | 0.67 | 1.00 | 0.85 | 247 |
| 7 | 0.00 | 1.00 | 0.01 | 2 | 0.00 | 1.00 | 0.02 | 2 | 0.01 | 1.00 | 0.02 | 2 | 1.00 | 0.50 | 0.71 | 2 |
| 8 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 |
| 9 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 |
| 10 | 0.00 | 0.00 | 0.00 | 13 | 0.00 | 0.00 | 0.01 | 13 | 0.05 | 1.00 | 0.09 | 13 | 0.05 | 1.00 | 0.10 | 13 |
| 11 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 |
| 12 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 | 0.00 | 0.00 | 0.00 | 0 |
| **mic. avg** | 0.18 | 0.54 | 0.27 | 649 | 0.17 | 0.53 | 0.29 | 649 | 0.31 | 0.91 | 0.47 | 649 | 0.32 | 0.92 | 0.50 | 649 |
| **mac. avg** | 0.16 | 0.49 | 0.19 | 649 | 0.15 | 0.49 | 0.20 | 649 | 0.49 | 0.88 | 0.49 | 649 | 0.50 | 0.88 | 0.52 | 649 |
| **w. avg** | 0.35 | 0.54 | 0.41 | 649 | 0.35 | 0.53 | 0.44 | 649 | 0.60 | 0.91 | 0.65 | 649 | 0.58 | 0.92 | 0.69 | 649 |
| **smp. avg** | 0.17 | 0.53 | 0.25 | 649 | 0.16 | 0.52 | 0.27 | 649 | 0.58 | 0.89 | 0.63 | 649 | 0.58 | 0.90 | 0.67 | 649 |

Table 13: Deployment — full classification reports (classes 0–12) for the four quantized deployed variants. L-3 is Llama-3 8B, and DS R1 is DeepSeek R1 7B. All metrics are represented from the deployment logs. Precision, Recall, F1-Score, and Support are represented as P, R, F1, and Sup. The cumulative results are represented in micro average, macro average, weighted average, and sample average.

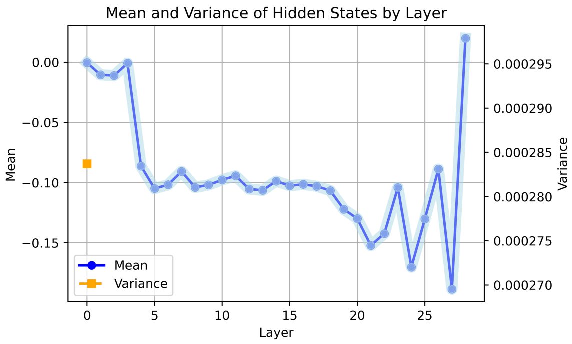


Figure 9: Mean and variance of hidden state activations across layers for the fine-tuned DeepSeek R1 Distill 7B. The x-axis represents the layer number, the left y-axis shows the mean of activations, and the right y-axis shows the average variance.

**Llama-3 8B Insights:** As depicted in Figure 10, the Llama-3 8B model exhibits a notable increase in the variance of its hidden states in the later layers (approximately layers 25-31). This region of increased variance is intriguingly correlated with the layers that demonstrate the highest probing performance for FoS classification (micro-F1 score > 0.60). The mean of the hidden states also displays a distinct pattern, with a discernible dip in these later layers. This suggests that as representations progress through Llama-3, they become more dispersed and potentially more separable for FoS-related features in the upper layers.

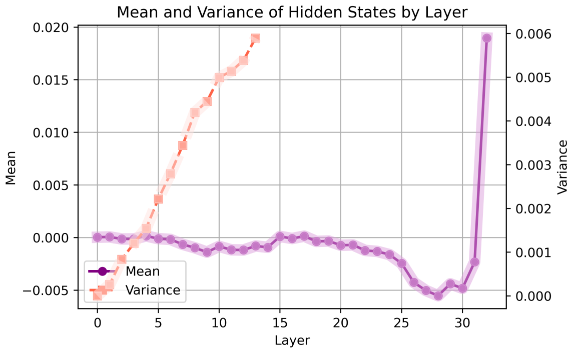


Figure 10: Mean and variance of hidden state activations across layers for the fine-tuned Llama-3 8B.

**Missclassification Analysis:** We further include a per-class misclassification analysis of the 16-bit quantized Llama-3 and DeepSeek R1 models. Figure 11 and Figure 12 show the relative counts of false positives and false negatives for each FoS label. These plots are derived from the fine-tuning of DeepSeek R1 and Llama-3, highlighting that skewed classes (e.g., frequent Alliteration, rare Apostrophe or Synecdoche) drive asymmetric error patterns. The plot reports per-class counts of false positives (Predicted=1, True=0) and false negatives (Predicted=0, True=1).

### A.11 Compute and Runtime

All experiments were conducted on NVIDIA A100 GPUs, CUDA 11.7, PyTorch 2.0, and HuggingFace Transformers v4.28. Full fine-tuning took ≈4.2 hours for DeepSeek R1 Distill 7B and ≈5.1 hours for Llama-3 8B per fold, using 14 GB and 20 GB of GPU memory, respectively.

### A.12 Additional Experimentation Details

Due to page limits, the following figures are provided here for reference:

**Fine-Tuning Loss Comparison.** Here we show in Figure 13 the comparative evaluation of the loss comparison among the quantized variations of Lllama-3 and DeepSeek R1 during fine-tuning.

**Fraction-wise Evaluation Loss.** Here we show in Figure 14 the final validation loss on the respective training subset size.

**Attention Heatmaps.** In accordance with Section 6.3, here we show one instance of the attention heatmaps of a sentence in Figure 15 and Figure 16, where both Llama-3 and DeepSeek R1 failed to identify the FoS labels.

The sentence here is: হৃদয়ের দ্বারে দ্বারে ভ্রমি মোর সারা নিশি প্রাণে প্রাণে খেলাইয়া প্রভাতে যাইব মিশি। (*Through the doors of every heart we roam all night, frolicking in each soul; at dawn we will go forth, blending as we go*). This sentence contains Personification (Label 3), Onomatopoeia (Label 4), and Alliteration (Label 6). However, both models mistake the FoS present as a Metaphor (Label 2). Similarly, for many other sentences, the internal layer representations could not decode the patterns present in the sentence.

### A.13 Discussion

Our extensive experiments reveal several key insights into fine-tuning and interpreting LLMs for Bengali FoS detection.

#### A.13.1 Performance Disparities Across FoS Labels

Analysis of per-label metrics (labels that were found and annotated for the Gold Standard dataset) in Table 14 with layer probing shows that frequent

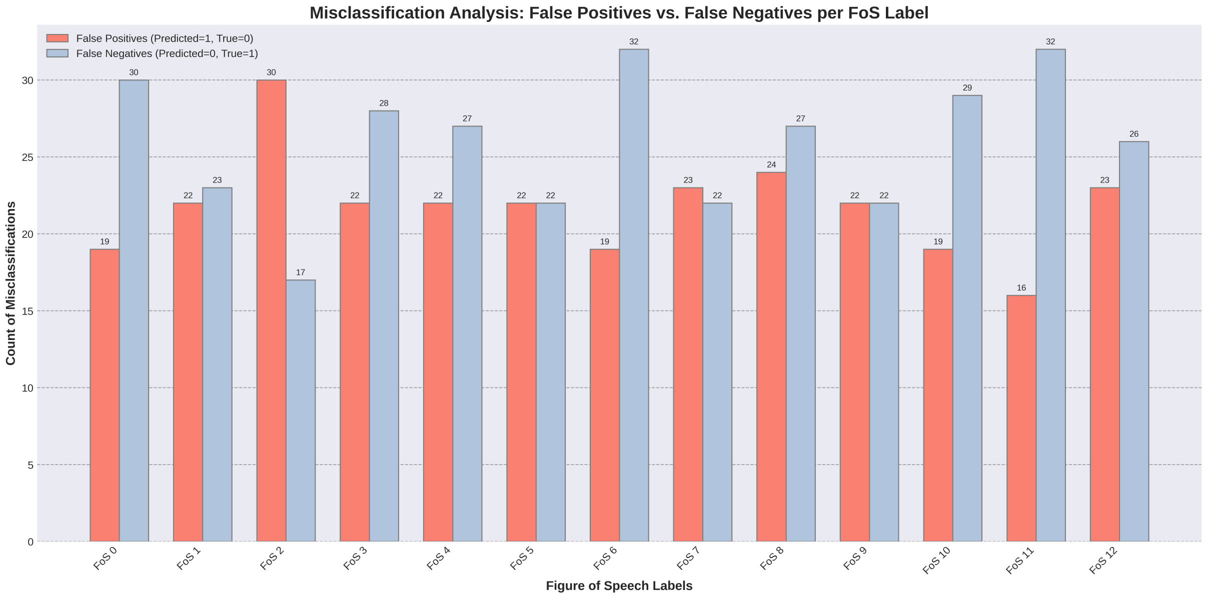


Figure 11: Misclassification analysis of the 16-bit quantized DeepSeek R1 Distill 7B model on BengFoS. Errors are unevenly distributed across labels, with frequent classes (e.g., Alliteration, None) exhibiting higher false positives and rare classes (e.g., Apostrophe, Synecdoche) showing higher false negatives.

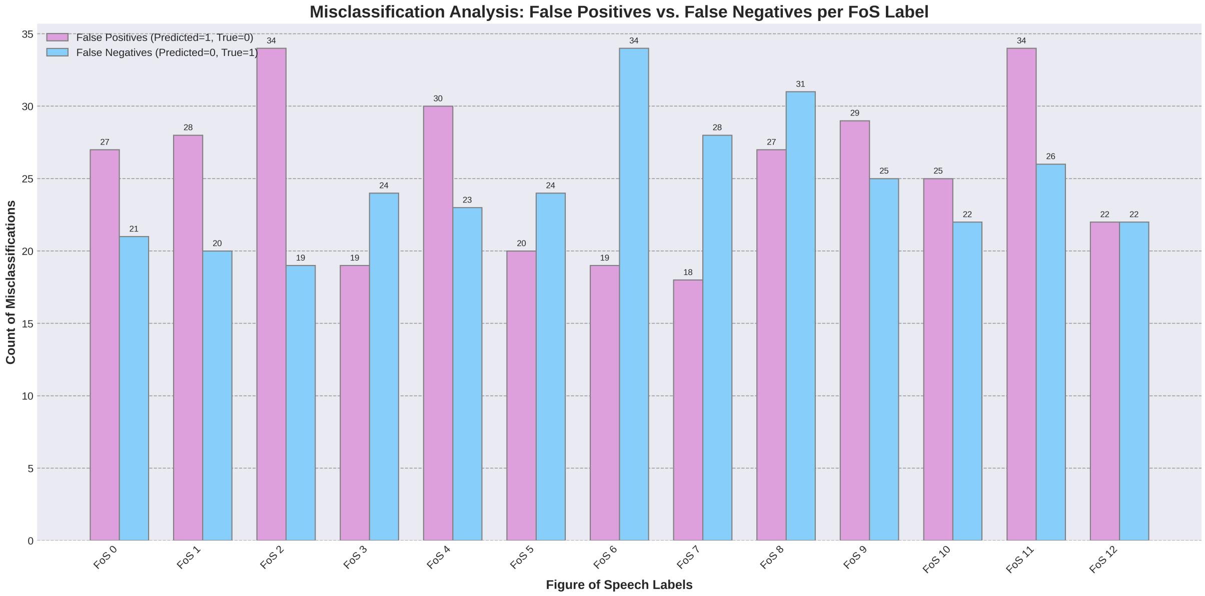


Figure 12: Misclassification analysis of the 16-bit quantized Llama-3 8B model on BengFoS. Compared to DeepSeek R1, Llama-3 exhibits higher false positives in several mid-frequency labels and persistent false negatives for rare labels. The visualization highlights the challenge of class imbalance and the asymmetric error patterns across different figures of speech.

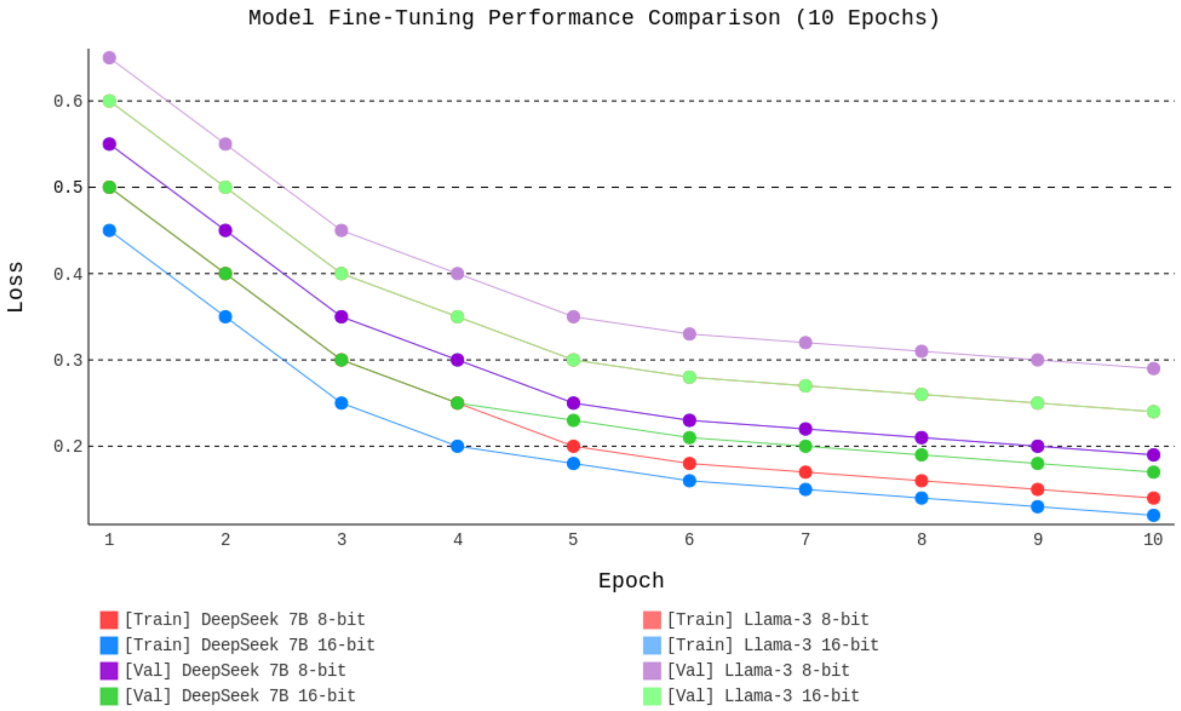


Figure 13: Loss comparison among the quantized variations of Lllama-3 and DeepSeek R1 during fine-tuning.

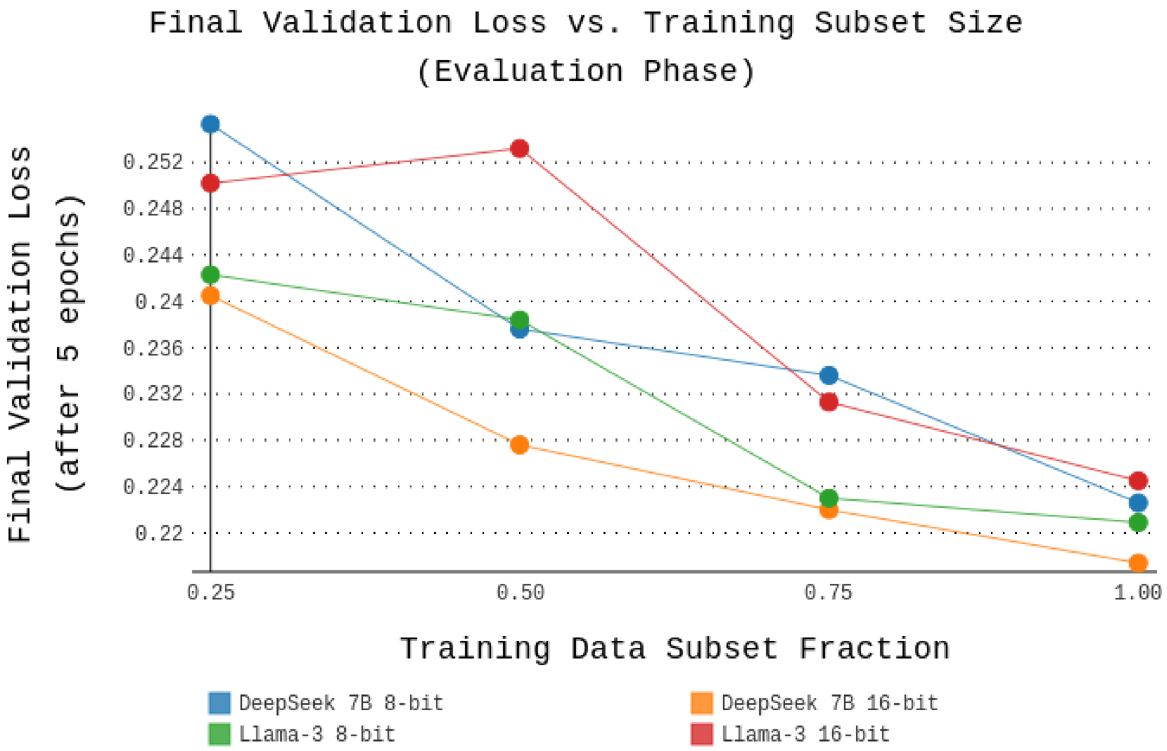


Figure 14: Fraction-wise validation loss of Llama-3 and DeepSeek R1

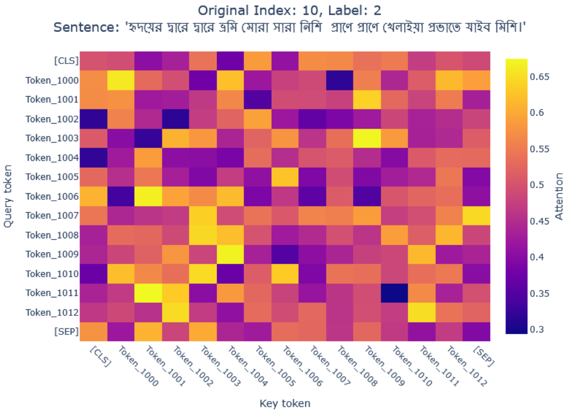


Figure 15: Attention heatmap generated by DeepSeek for a misclassified FoS identification sentence with multiple FoS labels.

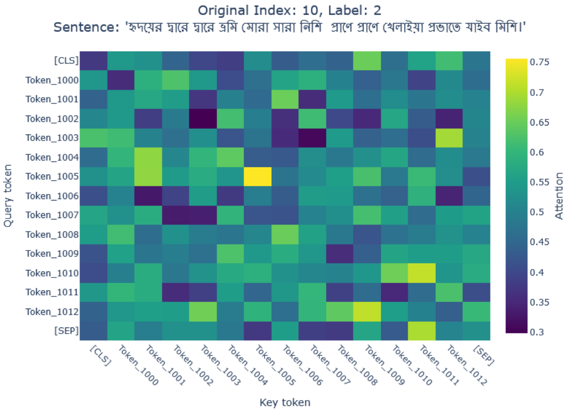


Figure 16: Attention heatmap generated by Llama for a misclassified FoS identification sentence with multiple FoS labels.

categories (e.g., None, Alliteration) achieve higher precision and recall, while rare categories (e.g., Apostrophe, Euphemism) suffer from near-zero recall. This imbalance underscores the need for data augmentation or cost-sensitive training to capture underrepresented FoS better. Also, though both models exhibit similar performances, Llama-3 was found to cover a better FoS Label range than DeepSeek R1.

### A.13.2 Impact of Deployment Quantization

Deployment experiments in Section 5.1 indicate that 16-bit quantization of Llama-3 and DeepSeek R1 incur minimal drops in macro-F1 (<2%), demonstrating the feasibility of real-time inference. However, rare classes saw larger degradation under quantization, hinting at reduced representation fidelity for nuanced FoS features.

## A.14 Fine-Grained Deployment Result Analysis

To assess practical applicability, we evaluated the 16-bit quantized fine-tuned DeepSeek R1 7B and Llama-3 8B models on the complete BengFoS test set. This evaluation aimed to simulate a deployment scenario and provide a comprehensive understanding of their classification capabilities for Bengali Figures of Speech (FoS).

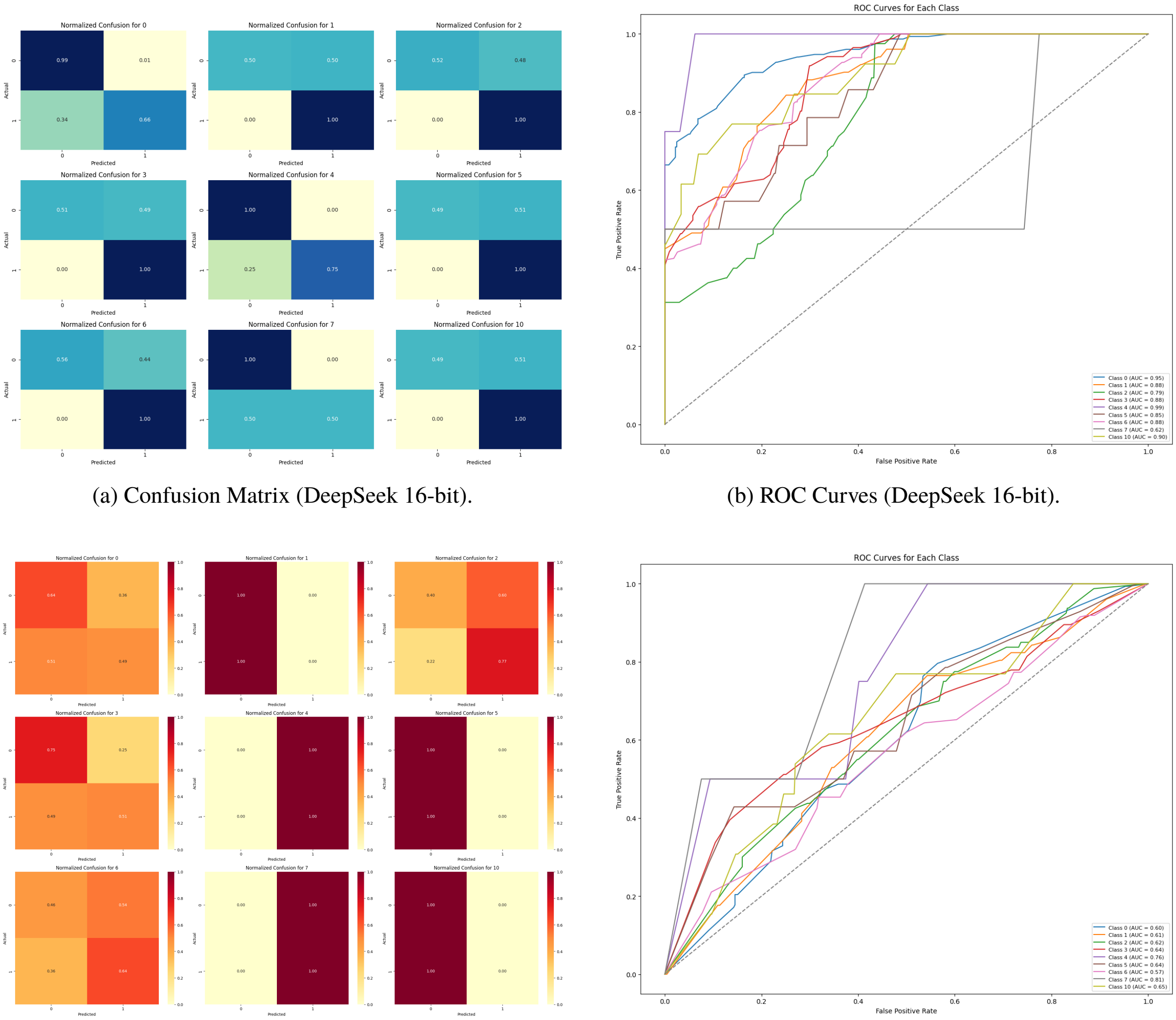
The DeepSeek R1 7B (16-bit) model demonstrated a weighted average F1-score of 0.64 and a macro average F1-score of 0.50. A detailed breakdown of its per-class performance and error patterns can be observed in its confusion matrix, presented in Figure 17a. The model's ability to discriminate between classes is further illustrated by the Receiver Operating Characteristic (ROC) curves and their corresponding Area Under the Curve (AUC) values, shown in Figure 17b. Notably, for several FoS classes, the DeepSeek R1 model achieved high AUC values (e.g., Class 0 AUC = 0.95, Class 4 AUC = 0.99), indicating strong discriminative power for these specific categories.

In comparison, the Llama-3 8B (16-bit) model achieved a weighted average F1-score of 0.40 and a macro average F1-score of 0.19. The confusion matrix in Figure 17c details its classification behavior across the different FoS labels, highlighting areas where it excels or struggles. The ROC curves, depicted in Figure 17d, provide a visual assessment of its class-wise true positive rate versus false positive rate. While Llama-3 showed reasonable AUCs for certain classes (e.g., Class 4 AUC=0.76, Class 7 AUC=0.81), its overall discriminative capability, as reflected by the average F1-scores and many per-class AUCs, was comparatively lower than DeepSeek R1 on this task.

These results suggest that while both models can be deployed at 16-bit precision, the DeepSeek R1 7B architecture, after fine-tuning, exhibits a more robust performance profile for Bengali FoS identification in this setting. The visual diagnostics (confusion matrices and ROC curves) in sub-figs. 17a to 17d are crucial for understanding the nuances of these performance differences beyond aggregate metrics.

## A.15 Error Analysis

Despite promising layer-wise trajectories, probing cannot fully reveal causal influence or disentan-



(a) Confusion Matrix (DeepSeek 16-bit).

(b) ROC Curves (DeepSeek 16-bit).

(c) Confusion Matrix (Llama-3 16-bit).

(d) ROC Curves (Llama-3 16-bit).

Figure 17: Performance evaluation of the 16-bit quantized versions of DeepSeek R1 7B and Llama-3 8B on the BengFoS test set. Subfigures (a) and (c) show the normalized confusion matrices, while (b) and (d) illustrate the per-class ROC curves and AUC values.

| | DeepSeek R1 Distill 7B | | | | Llama-3 8B | | | |
|---|---|---|---|---|---|---|---|---|
| FoS Label | Precison | Recall | F1-Score | Support | Precison | Recall | F1-Score | Support |
| FoS 0 | 0.276 | 0.973 | 0.429 | 147 | 0.397 | 0.211 | 0.276 | 147 |
| FoS 1 | 0.099 | 0.893 | 0.179 | 56 | 0.105 | 0.929 | 0.189 | 56 |
| FoS 2 | 0.152 | 0.716 | 0.251 | 81 | 0.152 | 0.963 | 0.263 | 81 |
| FoS 3 | 0.185 | 0.980 | 0.311 | 98 | 0.180 | 0.888 | 0.299 | 98 |
| FoS 4 | – | – | – | – | 0.006 | 1.000 | 0.011 | 3 |
| FoS 5 | 0.031 | 0.733 | 0.060 | 15 | 0.030 | 1.000 | 0.059 | 15 |
| FoS 6 | 0.438 | 0.456 | 0.447 | 239 | 0.449 | 0.146 | 0.221 | 239 |
| FoS 7 | – | – | – | – | 0.010 | 0.600 | 0.019 | 5 |
| FoS 9 | – | – | – | – | 0.003 | 1.000 | 0.006 | 1 |
| FoS 10 | 0.043 | 0.714 | 0.082 | 21 | 0.032 | 0.524 | 0.061 | 21 |

Table 14: Cumulative performance metrics for FoS labels during layer probing for both Fine-tuned Models. The probing analyses reveal that Llama-3 performed better in identifying slightly more diverged FoS labels (Labels 4, 7, and 9), while Llama-3 is more stable across the metrics.

gle distributed representations. Some FoS categories (e.g., Apostrophe, Euphemism) remain hard to probe due to their sparsity or dependence on discourse context. Inspection of misclassified or missing labels reveals common patterns as: **Semantic Overlap:** Sentences with multiple FoS labels (e.g., Metaphor + Personification) often confuse the model's single-head classifier. **Idiomatic Expressions:** Culturally specific idioms are frequently missed or misattributed. **Length Sensitivity:** Longer sentences with complex syntax yield lower accuracy, suggesting transformer truncation limits semantic capture.