

AND-2500: A Novel Assamese News Corpus and Comparative Benchmarking of Language Models

Pragyat Jyoti Baruah*, Arnab Paul, Vamanpally Sai Snehal, Tirthanka Borah

Assam University, Department of Computer Science and Engineering,
India

{pragyat.jyoti.baruah, arnab.paul, vamanpally.sai.snehal, tirthanka.borah}@aus.ac.in

Abstract. The rapid advancement of Large Language Models (LLMs) has transformed the landscape of Natural Language Processing (NLP), enabling remarkable progress in language understanding and generation. However, these advances remain unevenly distributed, particularly for low-resource languages such as Assamese, which suffer from a severe lack of high-quality open-sourced annotated corpora and standardized benchmarks. To address this gap, this paper introduces a novel AND-2500 dataset, which is manually curated by expert validation. The corpus comprises Assamese news articles annotated across 20 fine-grained categories following the International Press Telecommunications Council (IPTC) media topic taxonomy. This paper presents a comprehensive benchmarking framework that systematically evaluates both traditional encoder-only transformer models (MuRIL, IndicBERT, XLM-R etc.) and state-of-the-art generative LLMs (GPT2, Llama-3, Qwen2.5, Mistral and Aya-23) under supervised, zero-shot, and few-shot learning paradigms. Additionally, this paper assesses the generative capability of LLMs through a headline generation task, employing lexical and semantic evaluation metrics, including F1-score, BERTScore, and Semantic Answer Similarity (SAS). Our results demonstrate that supervised fine-tuned multilingual encoders like MuRIL remain the gold standard for classification (F1 score: 0.89) of Assamese text, whereas LLMs like Llama-3 can achieve near-supervised performance (F1: 0.88) with just twenty examples from AND-2500, offering a computationally efficient pathway for deploying LLMs in resource-constrained environments. This work establishes a vital baseline for Assamese NLP and promotes the open sourcing of linguistic resources for low-resource language.

Keywords. Dataset, LLMs, transformer language models, Assamese language, low-resource NLP.

1 Introduction

The field of Natural Language Processing (NLP) has witnessed a paradigm shift with the advent of transformer-based language models. These language models can be classified into two categories based on their design and objectives: (a) the encoder-only models (such as BERT, XLM, RoBERTa etc.) which are designed for understanding context, and (b) decoder-only Large Language Models (LLMs) like Llama, GPT2, GPT4 etc. which are designed for text generation. The main objective of the encoder-only models is to provide relevant contextual embeddings for a language. Whereas, decoder-only models mainly focus on down-stream tasks like next word prediction. However, whether the task is natural language understanding or natural language generation, the quality of these models remains heavily dependent on the usage of a large corpora.

Assamese is a classical Indian language [1] spoken by over 15 million people across India [2], primarily in the state of Assam. Despite its vast number of speakers, it lacks robust digital presence unlike high-resourced languages like English. Assamese, a low-resourced language, lacks annotated datasets and benchmarking suites, thus creating a research gap in this context.

To address this gap, this paper makes three primary contributions:

- **Dataset Curation:** The paper open-sources a news dataset which is curated and annotated. This dataset can be used across various NLP tasks of classification and generation. It can also contribute to the fine-tuning of LLMs.

- **Comparative Benchmarking:** Here, rigorous evaluation of the proposed dataset is performed using nine encoder-only transformer models and five Large Language Models for classification tasks. The performance of these models is compared using two additional standard corpora.
- **Generative Evaluation:** This paper explores the capability of LLMs to generate coherent Assamese text using prompting strategies on the dataset.

The paper is organized as follows: Section 1 provides the introduction of the study by clearly mentioning the contributions. Section 2 states the motivation of this study by highlighting the language characteristics and the research gaps. Section 3 gives a brief study of the related works taken for this study. Section 4 provides the description of the proposed AND-2500 dataset. Section 5 describes the methodological pipeline followed in this study. Section 6 gives the results and its discussion. Finally, the conclusion is drawn in section 7.

2 Motivation

2.1 Language Characteristics

Assamese is a linguistically rich inflectional language expressed by stringing together morphemes rather than using separate words [3]. It belongs to the Indo-European language family, which uses alphabets from the ‘Assamese script’ [4]. The Assamese script, derived from the Brahmi Script [5], consists of 11 vowels, 33 consonants, and 10 digits [6, 7]. Unlike English, it follows a complex writing style consisting of ‘conjunct consonants’ and ‘vowel diacritics’.

There are some multi-lingual generative AI models (like GPT5.2, DeepSeek etc.) that can generate domain-specific Assamese datasets. However, the necessity for curating a dedicated Assamese News Dataset (AND-2500) arises due to the unique linguistic characteristics, which most generic multilingual models often fail to capture:

- **Morphological Richness:** Assamese morphology allows for extensive inflection and derivation, which may lead to a high

vocabulary sparsity. A model trained on multilingual tokenizers may excessively fragment Assamese words into meaningless sub-word units, thus losing its semantic coherence.

- **Syntactic Flexibility:** While Assamese predominantly follows a Subject-Object-Verb (SOV) word order, it exhibits a degree of free word order due to its case-marking system. This structural divergence from the Subject-Verb-Object (SVO) order of English poses a significant challenge for transfer learning from English-centric generative models.
- **Script Specificity:** Assamese uses the ‘Assamese script’, which shares partial characteristics with the ‘Bengali/Bangla script’, but includes some unique characters (like ‘ৰ’ instead of ‘৳’, ‘৳’ and ‘৳’ instead of ‘৳’). Most generative AI models (like GPT) treat the scripts of Assamese and Bengali as identical and often hallucinate or misclassify these distinct characters, necessitating a dedicated curation of Assamese corpora.

2.2 Lack of Open-Sourced Dataset

Large language models require a large amount of fine-tuning using language dedicated corpus to produce efficient output. Assamese has a low digital presence on online platforms, as compared to other Indic languages like Hindi. Studies like [2] state that Assamese comprises of only 0.022% of Wikipedia. The current landscape of Assamese NLP resources is fragmented, with a small number of open-source corpora. Although some generative AI models can be used to create Assamese dataset, they need to undergo extensive validation by language-experts as these models fail to address every linguistic characteristic of the language.

3 Related Work

The journey of Indic NLP has progressed from early statistical tools to generative AI models. Initial efforts built foundational resources like POS taggers using rule-based methods. The discriminative era, marked by models like mBERT

[8], saw focused adaptations such as MuRIL [9] and IndicBERT [10] to better serve Indian languages. Now, the generative era which is led by models like GPT2 [11] and Gemma [12], offers remarkable capabilities but remains inadequately evaluated for low-resource languages like Assamese. This gap highlights the need for dedicated benchmarking to understand their true potential in local contexts.

In this context, various studies have explored the use of both encoder-only and decoder-only language models for Indic languages [2] including Assamese [13–17]. While encoder-only models like BERT (and its variants) are designed for providing better embeddings for Natural Language Understanding, the decoder-only models like GPT etc. are developed for Natural Language Generation tasks. The existing studies present in literature address the challenges of different NLP tasks for Indic languages across various language model architectures.

In [2], a comprehensive analysis of language models across various NLP tasks of Natural Language Understanding, and Natural Language Generation are studied for a set of Indic languages. The study states that LLMs are capable of handling Indic languages either because they are specifically pretrained and fine-tuned using Indic datasets, or they are pretrained on a multilingual dataset. It surveyed over four hundred large language models, and reported that only twenty-eight among those models were capable of handling Indic languages. Based on the findings, the paper categorized the twenty-eight models under two classes- Multilingual Large Language Models (MLLMs) and Indic language specific Large Language Models (ILLMs). The findings also state the problem of under-representation of Indic languages in standard pre-training corpora. It states that Indic languages like Hindi, Tamil and Urdu accounts for 0.256%, 0.265%, and 0.336% each of the total Wikipedia articles. Whereas, Assamese represents only 0.022% of Wikipedia articles and a common crawl data of 0.0023%. The study provides a comparative ranking of the performance of LLMs across twelve languages for the task of language understanding and language generation. While seven languages are categorized as 'high' performers, Assamese is classified as a 'medium' performer.

[13] created an Assamese corpus from eight diverse data sources like novel, biopic, news, and so on. The corpus has a collection of 752 line and 4,678 tokens. The study compared the performance of a rule-based model with a text-based question-prompt on LLM for the task of anaphora resolution for Assamese noun words. Their study reported a better performance by LLM over the rule-based approach used for the specified task.

[14] developed a Question Answering (QA) model for Assamese language. The study proposed a hybrid of Named Entity Recognition model, and transformer-based (XLM-R) word extraction and contextual embedding for creating the knowledge base. The ranking for answer retrieval is performed by using Cosine similarity on the word-pairs. They compared the performance of their proposed model with four pre-trained QA models- MuRIL, XLM-R, IndicBERT, BERT-base uncased. Their hybrid model was capable of outperforming only one model (i.e. IndicBERT). They created a QA dataset comprising a total of 1000 question-answer pairs.

[16] performed evaluation of three token encoders (used by LLMs)- Byte-Pair Encoding, WordPiece and Unigram for Assamese language. The study conducted a systematic evaluation using varying vocabulary sizes of 8000, 16000 and 32000 from an open-sourced Wikipedia dataset with 84137 unique tokens. The study found that Byte-Pair Encoding when evaluated with the maximum vocabulary size provided superior results. The findings also state that the size of the vocabulary is directly proportional to the quality of token-encodings provided by LLM tokenizers. Similarly, [17] performed intrinsic evaluation on five LLM tokenizers viz. gpt-4o_tokenizer, mistral_large_instruct_2407, meta_llama_405B, google_gemma_2_27b_it, sutra_mlt256_v2. Their findings state that sutra_mlt256_v2, which has the highest vocabulary size among the others performed best.

[18] introduces a large open-source Assamese dataset developed for neural machine translation. The complete corpus consists of both unlabeled as well as labeled datasets. The unlabeled data consists mostly of Wikipedia data, whereas the labeled data consists of ChatGPT-generated dataset, etc.

Table 1. A sample of the proposed AND-2500 dataset for readers' understanding (the url column is removed here)

Sl. No.	DoP (mm/dd/yyyy)	News source	Headline	News_body	Category
1	04/25/2021	News18 Assam	এতিয়া একাধিক স্মাৰ্টফোনত হোৱাটছএপ একাউণ্ট এটা ব্যৱহাৰ কৰিব পাৰিব, পৰিৱৰ্তনবোৰ সোনকালেই আহি আছে Transliteration: Etiyā ekādhik smārtpḥōnata hōyāts'ēpa ēkā'unṭa ēṭa byabahāra kariba pāriba, pariṛartanabōra sōnakālē'i āhi āchē Translation: One can now use a single WhatsApp account on multiple smartphones, the changes are arriving soon	এতিয়াৰ পৰা, একে সময়তে একাধিক স্মাৰ্টফোনত হোৱাটছএপৰ একাউণ্ট খুলিব পাৰি। হোৱাটছএপৰ মাল্টি-ডিভাইচ সামঞ্জস্যতা সুবিধা সোনকালেই মুকলি কৰা হ'ব।.....ইয়াত অন্তৰ্ভুক্ত আছে ইমোজি প্ৰতিক্ৰিয়া, ২জিবি ফাইল ভাগ বতৰা কৰা আৰু গ্ৰুপ ভইচ কলত ৩২ জন লোক যোগ দিয়াৰ দৰে সুবিধা। Transliteration: Etiyāra paraṃ, ēkē samayātē ēkādhik smārtpḥōnata hōyāts'ēpara ēkā'unṭa khuliba pāri. Hōyāts'ēpara mālti-dibhā'isa sāmāñjasyatā subidhā sōnakālē'i mukali karā ha'ba.....Iyāta antarbhuṭta āchē Imōji pratikriyā, 2 jibi phā'ila bhāga batara karā ārau grupa bho'isa kalata 32 jana lōka yōga diyāra darē subidhā. Translation: From now on, one can use a single WhatsApp account across multiple smartphones simultaneously. WhatsApp's multi-device compatibility feature will be rolled out soon.....These include emoji reactions, sharing files upto 2 GB, and enabling upto 32 participants in group voice calls.	Technology

[5] developed an Assamese named entity dataset comprising 99000 tokens. The dataset was evaluated using various language models, where MuRIL performed best.

[19] created a parts of speech dataset for Assamese language. This dataset comprises of 283506 tokens that are mapped across 20280 sentences.

4 Dataset Description

4.1 Collection

The AND-2500 dataset [20] is prepared by manual collection from various popular newspapers of Assam. Each news article is carefully reviewed and extracted from its source by human curators, to preserve the linguistic authenticity and contextual accuracy of the text.

This intensive method is deliberately used instead of automatic web-scraping, so as to

produce a validated and reliable high-quality corpus.

The primary sources of data collection are: ePapers viz. Dainik Janambhumi (দৈনিক জনমভূমি) and Amar Asom (আমাৰ অসম), and four online news portals viz. Asomia Pratidin (অসমীয়া প্ৰতিদিন), Dainandin Barta (দৈনন্দিন বাৰ্তা), Niyomia Barta (নিয়মীয়া বাৰ্তা), All India Radio, Dainik Agradoot (দৈনিক অগ্ৰদূত) and News18 Assam.

The primary dataset consists of the following attributes: Date of Publication (DoP), news source, headline, news body and the url link of the website articles. The complete dataset consists of 2500 news articles in Assamese script. Table 1 displays a sample of the AND-2500 with added translation and transliteration (ISO 15919 style) for reader's understanding.

4.2 Taxonomy and Annotation

The collected articles are annotated by strictly following a global standard IPTC media topic

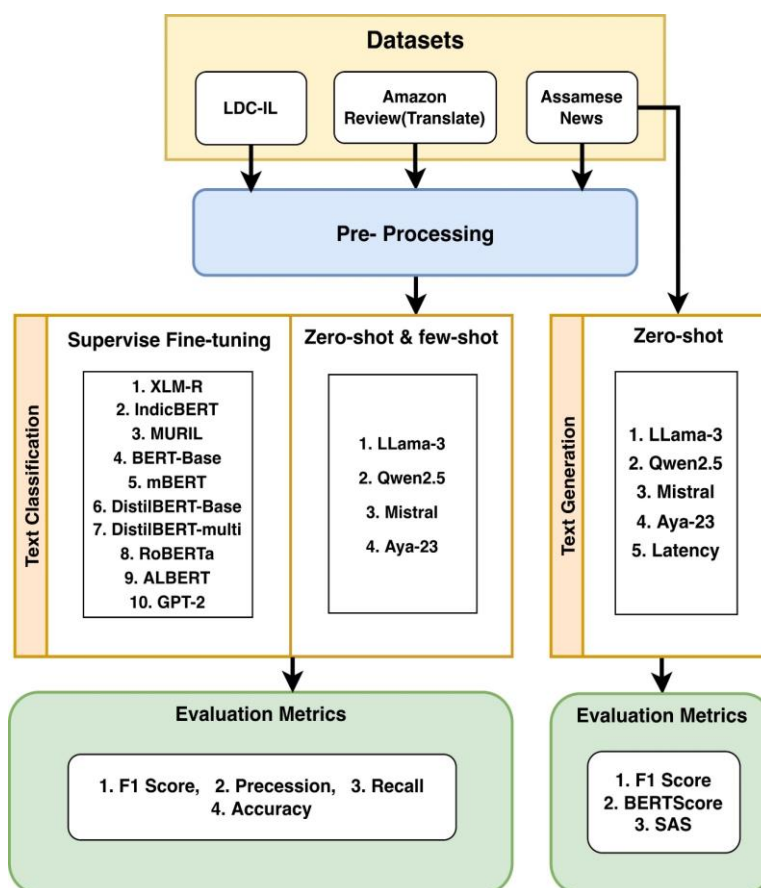


Fig. 1. The methodology pipeline for benchmarking AND-2500

taxonomy [21]. The IPTC [22] stands for International Press Telecommunications Council defines a controlled vocabulary for news topics or categories.

The 2500 articles are annotated into twenty distinct news categories/labels by following the defined taxonomy: ‘Accident’, ‘Agriculture’, ‘Architecture’, ‘Crime, law & Justice’, ‘Culture’, ‘Disaster’, ‘Economy, business & finance’, ‘Entertainment’, ‘Examination’, ‘Festival’, ‘Health’, ‘Literature’, ‘Lifestyle’, ‘Politics’, ‘Social Issue’, ‘Social Media’, ‘Sports’, ‘Technology’, ‘Weather’, and ‘Wildlife’.

The authors acknowledge that the annotated corpus comprises of some news labels which might have fine distinctions of semantic proximity. However, the authors chose to maintain a high degree of granularity to ensure the novelty and depth, thus aligning to broader future goals.

5 Methodology

To evaluate the efficiency of our dataset, we designed two experimental frameworks: Text Classification and Text Generation.

Two additional datasets are cross-evaluated for text classification to compare them with our dataset. Fig. 1 presents an overview of the methodology.

5.1 Model Selection

To ensure that our work remains accessible and reproducible for all researchers, only open-sourced language models are considered in this study. The selection process further required that the model demonstrate capability in processing and generating Assamese text.

Table 2. Number of tokens of the complete datasets

Dataset	No. of tokens
AND-2500	259177
LDC-IL	10127030
Amazon reviews (translated)	89266

This paper considers two sets of models- BERT-based (as encoder-only language model), and generative AI large language model (decoder-only language model). In the study, nine BERT-based language models are found compatible for Assamese, whereas four LLMs are found to be compatible for Assamese text generation. The GPT2 model (LLM) is found incompatible for Assamese text generation.

However, it is used as a supervised baseline by fine-tuning the model with our dataset for text classification.

The list of the BERT-based models selected:

- Indic-Focused: MuRIL [9], IndicBERT [10].
- Multilingual-setup: mBERT [8], XLM-R [23], DistilBERT-Multilingual [24].
- General-models: BERT-Base [25], RoBERTa [26], ALBERT [27], DistilBERT-Base [28].
- The list of LLMs selected:
- For supervised text classification: GPT2 [11].
- For text generation: Meta-Llama-3-8B-Instruct (termed as Llama-3) [29], Qwen2.5-7B-Instruct (termed as Qwen2.5) [30], Mistral-7B-Instruct-v0.3 (termed as Mistral) [31], and aya-23-8B (termed as Aya-23) [32]. The GPT2 was excluded as it failed to produce coherent output without training.

Llama-3 is the latest dense LLM model provided by Meta, with a parameter size of 8 billion. The Qwen2.5 with 7 billion parameters is developed by Alibaba and is known for its strong multilingual reasoning. Mistral is trained using 7 billion parameters and is a highly efficient model that uses sliding window attention. Aya-23, trained using 8 billion parameters is a model by Cohere, specifically fine-tuned for 23 languages.

5.2 Experimentation Data and Protocols

The following datasets are used in our study:

- **LDC-IL dataset [33]:** This dataset is a gold-standard Assamese raw corpus validated and provided by the Linguistic Data Consortium for Indian Languages, Govt. of India. The complete dataset consists of texts that are categorized under six domains/labels- Aesthetics, Commerce, Mass Media, Official Document, Science & Technology, and Social Sciences. Among these labels, the four labels (i.e. Aesthetics, Mass Media, Science & Technology, and Social Sciences) consists of 99.33% of the total dataset. This study considers only these four labels for further experimentation. The experimental data retrieved from this corpus contains 9963620 tokens.
- **Amazon reviews (translated) dataset:** The Amazon reviews [34], is an open-sourced English corpus which contains reviews from 6643669 users on 2441053 products. For our experiment, only two columns were selected from the raw-dataset: Review-text, and rating. Amazon provides a rating range of 1 to 5. These ratings were converted into three sentiment labels- positive (rating 4 and 5), neutral (rating 3), and negative (rating 1 and 2). The review-texts are translated into Assamese using Microsoft Translator. The translated dataset is balanced by scaling and concatenation for its further evaluation.
- **AND-2500 dataset [20]:** For experimentation, only five labels were considered from the dataset- 'Sports', 'Entertainment', 'Health', 'Politics', 'Crime, law & justice'. This scaling-down strategy is applied to enable consistent and fair comparative analysis with the other datasets used for this study. The selected labels were not only chosen for this practical alignment, but also because they collectively represent a substantial and well-balanced portion, i.e. 54.44% of the original dataset.

The training protocols used in the study are as follows:

- **Supervised Fine-Tuning:** This method uses 70% of the total corpus for training the model,

Table 3. Results of the models after supervised fine-tuning

Model	AND-2500 dataset				LDC-IL dataset				Amazon reviews (translated) dataset			
	F1	Recall	P	Acc (%)	F1	Recall	P	Acc (%)	F1	Recall	P	Acc (%)
MuRIL	0.8939	0.8913	0.8991	89.24	0.6929	0.6773	0.7139	81.62	0.9605	0.9609	0.9622	96.09
XLM-R	0.8487	0.8519	0.8509	84.60	0.6966	0.6872	0.7098	80.03	0.8690	0.8697	0.8743	86.97
IndicBERT	0.2265	0.3306	0.1838	40.10	0.6114	0.5969	0.6333	76.16	0.6595	0.6717	0.6933	67.10
BERT-Base	0.2586	0.3325	0.2367	38.63	0.5527	0.5377	0.6155	74.74	0.7383	0.7394	0.7391	73.94
mBERT	0.3916	0.4473	0.4558	50.86	0.5182	0.5160	0.5214	76.93	0.8763	0.8762	0.8777	87.62
DistilBERT-Base	0.7897	0.7997	0.8040	78.24	0.6731	0.6845	0.6647	78.24	0.5997	0.6286	0.6538	62.86
DistilBERT-Multi	0.7897	0.7997	0.8040	78.24	0.7897	0.7997	0.8040	78.24	0.8245	0.8241	0.8254	82.41
RoBERTa	0.0823	0.2000	0.0518	25.92	0.3685	0.4024	0.3418	67.78	0.3826	0.4788	0.3301	47.88
ALBERT	0.1690	0.2651	0.1334	31.05	0.1509	0.2500	0.1081	43.23	0.1714	0.3387	0.1147	33.87
GPT2	0.4147	0.4341	0.4436	71.03	0.2336	0.4036	0.3219	40.36	0.6601	0.6612	0.6625	66.12

15% for validation, and 15% for testing. This method is followed for the text classification task in the study.

- **Zero-Shot:** Here, the LLM models were provided with a prompt with an input text sample to generate the desired output. This approach is tested on 10 samples per label for each of the dataset. In our experiment, zero-shot of the models is performed for both the text classification and text generation tasks.
- **Few-Shot:** This strategy provides a small set of test samples to the pre-trained LLMs, to perform the desired task on a new input text sample. This strategy is followed for the text classification task. In this study, 20 different samples for each label are provided to the LLMs to assign labels for 10 text samples (without any repetition from the examples given prior) for each label.

Table 2 displays the number of tokens present in the datasets used for this study.

5.3 Data Processing and Interpretation

5.3.1 Data Cleaning

The selected datasets are pre-processed before feeding them to the language models. The pre-processing pipeline implemented to prepare the data for experimentation is as follows:

- **Tokenization:** The input texts are segmented into word-level tokens based on whitespace to identify token boundaries.
- **Unwanted character removal:** Here, characters like emoticons, url, email addresses etc. are removed. The resulting cleaned dataset consists of Assamese words (in Assamese script) and alphanumeric characters.
- **Stop-word removal:** A list of common Assamese stop-words is selected for removal from the text classification dataset. It is used to remove high-frequency, low-information

Table 4. Results of zero-shot and few-shot using LLMs for text classification

Model	Parameters(in billion)	Our dataset				LDC-IL dataset				Amazon reviews (translated)			
		F1	Recall	P	Acc (%)	F1	Recall	P	Acc (%)	F1	Recall	P	Acc (%)
Llama-3 (zero-shot)	8	0.8600	0.870	0.8910	87.00	0.4190	0.4250	0.4620	42.50	0.8320	0.8330	0.8470	83.33
Llama-3 (few-shot)		0.8860	0.8900	0.9130	89.00	0.5320	0.5370	0.5800	53.75	0.8150	0.8170	0.8320	81.67
Qwen2.5 (zero-shot)	7	0.8020	0.8100	0.8140	81.00	0.4020	0.3600	0.6610	36.00	0.7130	0.7170	0.7190	71.67
Qwen2.5 (few-shot)		0.8640	0.8700	0.8870	87.00	0.5780	0.5750	0.5880	57.50	0.7990	0.8000	0.8010	80.00
Mistral (zero-shot)	7	0.5990	0.6100	0.7410	61.00	0.2450	0.3120	0.4050	31.25	0.6730	0.6830	0.7080	68.33
Mistral (few-shot)		0.8000	0.8000	0.8120	80.00	0.4060	0.4050	0.4930	40.54	0.6930	0.7000	0.7390	70.00
Aya-23 (zero-shot)	8	0.7550	0.7600	0.8130	76.00	0.4100	0.4000	0.4210	42.50	0.5870	0.6170	0.7460	61.67
Aya-23 (few-shot)		0.8200	0.8300	0.8460	83.00	0.5290	0.5370	0.6850	53.75	0.6360	0.6670	0.7540	66.67

words, reducing feature sparsity for the models for the task of classification.

5.3.2 Evaluation Metrics

For the task of evaluating the text classification task, four metrics are used:

- Precision (or P) measures the performance of a model by calculating the ratio of true positives with the total number of positive predictions during testing.
- Recall calculates the true positive rate of the predicted labels during testing.
- F1 Score (i.e. the harmonic mean of precision and recall) is used to account for any potential class imbalances.
- Accuracy (or Acc) is the overall percentage of correct classifications during testing.

For evaluating the text generation outputs, three metrics are used:

- **F1 Score:** The generated texts are matched with the ground-truth to measures the exact word overlap (n-grams).
- **BERTScore [35]:** This metric measures the semantic similarity of the generated text by

comparing it with the ground-truth answer. It uses pre-trained embeddings to measure semantic similarity. Here, the MuRIL is chosen for evaluating the BERTScore, as it is pre-trained in Assamese and provides a strong Assamese context embedding.

- **SAS (Semantic Answer Similarity):** It uses cross-encoders to determine if the generated output conveys the same meaning as the ground truth. It uses cosine similarity of the embeddings to evaluate the score.
- **Efficiency:** Average Inference Latency (in seconds) per generation is calculated for the models.

5.4 Experimentation Environment

All the experimentation of the models is conducted on the Kaggle Notebooks (open-sourced).

The hardware configuration of the cloud-based environment is: CPU: Intel(R) Xeon(R) @ 2.00GHz (4 Cores); GPU: 2 × NVIDIA Tesla T4 (16 GB VRAM each, 32 GB total); RAM: 30 GB.

Table 5. Text generation results using zero-shot of LLMs on the AND-2500 dataset

Model	Parameter (in billions)	Metric	AND-2500 dataset				
			Sports	Health	Politics	Entertainment	Crime, law & justice
Llama-3	8	F1	0.2185	0.1284	0.1182	0.1131	0.0945
		BERTScore	0.5238	0.3517	0.3994	0.4272	0.4821
		SAS	0.9577	0.8179	0.9209	0.9505	0.9691
Qwen2.5	7	F1	0.1645	0.1178	0.1498	0.2439	0.1107
		BERTScore	0.4664	0.4595	0.4518	0.5629	0.5578
		SAS	0.9462	0.9637	0.9601	0.9646	0.9706
Mistral	7	F1	0.1541	0.1466	0.1311	0.2405	0.1787
		BERTScore	0.4372	0.4402	0.4198	0.5294	0.4694
		SAS	0.9195	0.8423	0.9581	0.9473	0.8885
Aya-23	8	F1	0.2383	0.1582	0.1531	0.2194	0.0740
		BERTScore	0.5058	0.4370	0.4734	0.5436	0.4722
		SAS	0.9330	0.8798	0.9570	0.9507	0.9507

6 Results and Discussions

6.1 Classification Task

6.1.1 Supervised Fine-Tuning with Dataset

The results of the supervised experiments (in Table 3) reveal a distinct hierarchy of model suitability for Assamese. The MuRIL transformer outperformed the remaining 9 models for the task of text-classification achieving the highest accuracy scores for the three datasets. The model is specifically pretrained in seventeen Indic languages including Assamese.

This allows the model to represent Assamese words as whole tokens rather than some fragmented sub-words, thus creating superior semantic embeddings for them. Models like IndicBERT, mBERT, XLM-R, DistilBERT-multi are pretrained in Assamese and produced satisfactory results. The RoBERTa and ALBERT models performed poorly as these models are English-centric models. Although these models processed the text data, they were unable to provide strong embedding for Assamese. In contrast, BERT-Base

and DistilBERT-Base are not pre-trained in Assamese, but achieved higher accuracy score as they are able to provide semantic embeddings from the fine-tuning dataset.

The GPT2, a LLM, achieved moderate F1 scores when fine-tuned with the datasets. However, it required the heavy lifting of weight updates to learn the language patterns from scratch. Unlike MuRIL, it possessed no prior knowledge of Assamese, making it data-inefficient. Furthermore, the computational cost for supervised fine-tuning of LLMs is expensive and time-consuming. So, this study limits only the use of GPT2 model as the baseline.

6.1.2 Zero and Few-Shot of LLMs

Table 4 presents the performance of LLMs for the task of text classification in zero-shot and few-shot settings. Llama-3 demonstrated exceptional few-shot performance, achieving an F1 score of 0.8860 on the AND-2500. It also achieved the best accuracy in zero-shot setting for the proposed dataset. The transition from zero-shot to few-shot yielded massive improvements. Notably, Mistral

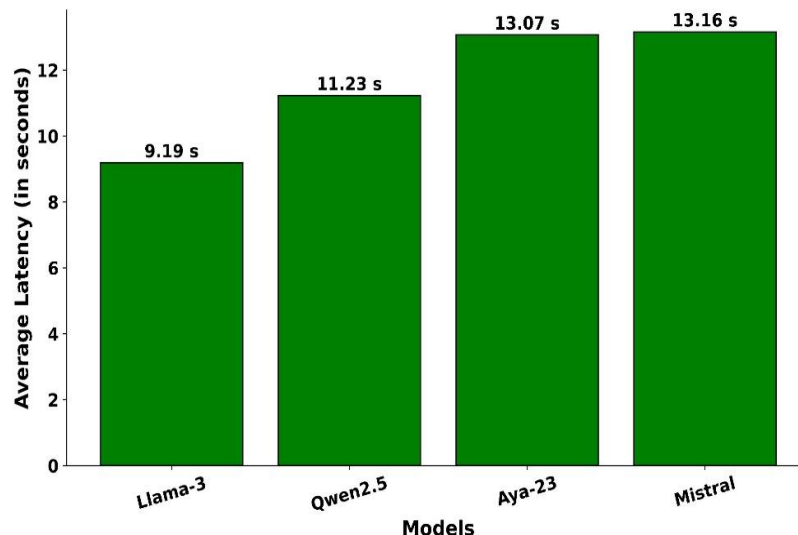


Fig. 2. Average latency of the LLMs for text generation

saw its F1 score jump from 0.5990 (zero-shot) to 0.8000 (few-shot).

This suggests that the LLMs provide effective text classification during few-shot when compared with zero-shot. All the models performed better on our dataset.

6.2 Text Generation

This task served as a validation of the content quality of the proposed AND-2500 dataset. Here, four LLMs are tasked for generating a headline based on the new content using zero-shot strategy. The generated headlines are then checked for lexical similarity using F1 score, whereas the semantic similarity is checked using BERTScore and SAS (in Table 5).

The F1 scores for generated headlines were low (between 0.07 and 0.25). The BERTScore is satisfactory with most scores ranging between 0.40 to 0.56. However, the SAS scores were extremely high, consistently exceeding 0.90. Qwen2.5 achieved a near-perfect SAS of 0.9706 in the Crime category. This confirms that the models were able to understand the context of the news bodies in our dataset. Due to its morphological richness, Assamese often expresses similar meanings with varied word forms. Consequently, the generated headline may be semantically

correct but lexically different from the ground truth, leading to low scores in lexical evaluation metric like F1 score. Qwen2.5 outperformed the other models for the task of text generation. The data and code used for this task are available in [36].

Fig. 2 shows the average latency of the models during text generation. In terms of latency, Llama-3 was the most efficient (9.19s per generation), followed by Qwen2.5 (11.23s). Mistral and Aya-23 were significantly slower (>13s).

6.3 Discussion

The high performance of both supervised (MuRIL) and few-shot (Llama-3) models confirms that the proposed AND-2500 dataset is a clean and solvable NLP dataset.

The strict data collection and systematic annotation pipeline were followed during data collection to avoid a noisy and unvalidated web-scraped dataset.

Conventionally, the standard pipeline for low-resource NLP involved pre-training and fine-tuning of encoder-only models like BERT and its variants. This methodology is computationally expensive when compared to LLM prompting.

Our results show that by simply prompting LLMs in zero-shot and few-shot settings yielded

comparable results. This indicates a distinct trend-shift of low-resource NLP towards LLMs.

7 Conclusion

This research presented a holistic effort to bridge the digital divide for the Assamese language and LLMs for Assamese.

The paper introduced the AND-2500 dataset, a standardized and open-source corpus that addresses the critical lack of multi-domain resources for NLP in low-resource languages.

Our comprehensive benchmarking established MuRIL as the current gold standard for supervised classification tasks in Assamese, offering high reliability where training data is available. Simultaneously, the potential of LLMs is also discussed, with Llama-3 achieving high accuracy (89%) in few-shot prompting for text classification task. Furthermore, the high Semantic Answer Similarity (SAS) scores in our headline generation task validate the richness of our dataset, proving it to be a reliable benchmark for future summarization and generation tasks.

References

1. **Kumar, S., Kataria, R., Nair, A. (2024).** Status of Classical Language: An Explainer, Research Unit, Press Information Bureau. Government of India (Ministry of Culture).
2. **Vaidya, A., Prabhakar, T., George, D. (2025).** Analysis of Indic Language Capabilities in LLMs. arXiv. DOI: 10.48550/arxiv.2501.13912.
3. **Saharia, N., Konwar, K.M., Sharma, U. (2013).** An Improved Stemming Approach Using HMM for a Highly Inflectional Language. Computational Linguistics and Intelligent Text Processing, pp. 164–173. DOI: 10.1007/978-3-642-37247-6_14.
4. **Verma, T.P. (1976).** Development of Script in Ancient Kamrupa, Asam Sahitya Sabha. Chandrakanta Handiqui Bhawan.
5. **Pathak, D., Nandi, S., Sarmah, P. (2022).** AsNER - Annotated Dataset and Baseline for Assamese Named Entity Recognition. Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6571–6577.
6. **Deka, D.S. (2006).** Rachana Bichitra, Assam Book Depot. Pan Bazaar.
7. **Sharma, P., Sharma, U., Kalita, J. (2012).** Suffix Stripping Based NER in Assamese for Location Names. 2012 2nd National Conference on Computational Intelligence and Signal Processing (CISP), pp. 91–94. DOI: 10.1109/nccisp.2012.6189684.
8. **Libovický, J., Rosa, R., Fraser, A. (2019).** How Language-Neutral Is Multilingual BERT?. arXiv. DOI: 10.48550/arxiv.1911.03310.
9. **Khanuja, S., Bansal, D., Mehtani, S. (2021).** MuRIL: Multilingual Representations for Indian Languages. arXiv. DOI: 10.48550/arxiv.2103.10730.
10. **Kakwani, D., Kunchukuttan, A., Golla, S. (2020).** IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-Trained Multilingual Language Models for Indian Languages. Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961.
11. **Radford, A., Wu, J., Child, R. (2019).** Language Models Are Unsupervised Multitask Learners.
12. **Gemma, T., Google, D. (2024).** Gemma: Open Models Based on Gemini Research and Technology. arXiv. DOI: 10.48550/arxiv.2403.08295.
13. **Das, M., Senapati, A., Senapati, A. (2025).** Prompt-Based Anaphora Resolution in Large Language Models. Computacion y Sistemas, Vol. 29, No. 3, pp. 1271–1281. DOI: 10.13053/cys-29-3-5917.
14. **Lahkar, M.P., Sharma, U., Pradhan, T. (2025).** When Graphs Meet Fine-Tuning: A Dual Approach to Low-Resource Assamese QA. 2025 6th International Conference on Recent Advances in Information Technology (RAIT), pp. 1–6. DOI: 10.1109/rait65068.2025.11089369.
15. **Nath, B., Pandey, K.K., Yadav, V. (2025).** Few-Shot and Zero-Shot Assamese Hate Speech Detection: A Comparative Benchmark of Large Language Models. Social Network

- Analysis and Mining, Vol. 16, No. 14. DOI: 10.1007/s13278-025-01531-x.
16. **Nath, B., Tamang, S. (2025).** Language-Specific Tokenization for Assamese: Efficiency and Downstream Integration with LLMs. *Impact in Computics*, Vol. 1, pp. 1–10. DOI: 10.65500/computics-2025-004.
 17. **Tamang, S., Bora, D.J. (2025).** Performance Evaluation of Tokenizers in Large Language Models for the Assamese Language. *International Journal of Information Technology*, Vol. 17, No. 4, pp. 2329–2332. DOI: 10.1007/s41870-025-02454-8.
 18. **Tamang, S., Bora, D.J. (2024).** Enhancing Assamese NLP Capabilities: Introducing a Centralized Dataset Repository. *arXiv*. DOI: 10.48550/arxiv.2410.11291.
 19. **Sarma, S.K., Talukdar, K. (2024).** Enabling Natural Language Processing and AI Research in Low-Resource Languages: Development and Description of an Assamese UPoS Tagged Dataset. *Journal of Electrical Systems*, Vol. 20, No. 3, pp. 1312–1320. DOI: 10.52783/jes.1506.
 20. **AUnlpLAB. (2025).** AND-2500-Dataset. GitHub.
 21. **I.P.T.C. (2022).** News Categories Taxonomy for the Media. IPTC Information Technology for News.
 22. **I.P.T.C. (2022).** International Press Telecommunication Council. IPTC Information Technology for News.
 23. **Conneau, A., Khandelwal, K., Goyal, N. (2019).** Unsupervised Cross-Lingual Representation Learning at Scale. *arXiv*. DOI: 10.48550/arxiv.1911.02116.
 24. **Yadagiri, A., Mohana, R., Pakray, P. (2025).** CNLP-NITS-PP at GenAI Detection Task 2: Leveraging DistilBERT and XLM-RoBERTa for Multilingual AI-Generated Text Detection. *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pp. 307–311.
 25. **Devlin, J., Chang, M.W., Lee, K. (2018).** BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. DOI: 10.48550/arxiv.1810.04805.
 26. **Liu, Y., Ott, M., Goyal, N. (2019).** RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*. DOI: 10.48550/arxiv.1907.11692.
 27. **Lan, Z., Chen, M., Goodman, S. (2019).** ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *arXiv*. DOI: 10.48550/arxiv.1909.11942.
 28. **Sanh, V., Debut, L., Chaumond, J. (2019).** DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv*. DOI: 10.48550/arxiv.1910.01108.
 29. **Touvron, H., Lavril, T., Izacard, G. (2023).** LLaMA: Open and Efficient Foundation Language Models. *arXiv*. DOI: 10.48550/arxiv.2302.13971.
 30. **Bai, J., Bai, S., Chu, Y. (2023).** Qwen Technical Report. *arXiv*. DOI: 10.48550/arxiv.2309.16609.
 31. **Jiang, A.Q., Sablayrolles, A., Mensch, A. (2023).** Mistral 7B. *arXiv*. DOI: 10.48550/arxiv.2310.06825.
 32. **Aryabumi, V., Dang, J., Talupuru, D. (2024).** Aya 23: Open Weight Releases to Further Multilingual Progress. *arXiv*. DOI: 10.48550/arxiv.2405.15032.
 33. **Ramamoorthy, L., Choudhary, N.K., Sharma, A. (2021).** A Gold Standard Assamese Raw Text Corpus. Central Institute of Indian Languages, Mysore.
 34. **Jain, K. (2021).** Amazon Reviews. Kaggle.
 35. **Zhang, T., Kishore, V., Wu, F. (2019).** BERTScore: Evaluating Text Generation with BERT. *arXiv*. DOI: 10.48550/arxiv.1904.09675.
 36. **AUnlpLAB (2024).** Zero-Shot-Text-Generation. GitHub.

Article received on 06/12/2025; accepted on 16/01/2026.

**Corresponding author is Pragyat Jyoti Baruah.*