Meaning Over Morphology: A Multi-Metric Benchmark of LLMs for Bangla Dialect Translation

Soumik Deb Niloy[1], **Subhey Sadi Rahman**[2], **Mahbub E Sobhani**[1,2], **Golam Rabiul Alam**[1], **Farig Sadeque**[1], **Md. Rezuwan Hassan**[1]

[1]BRAC University, Dhaka, Bangladesh, [2]United International University, Dhaka, Bangladesh

msobhani2410011@mscse.uiu.ac.bd | md.rezuwan.hassan@g.bracu.ac.bd

Abstract

Regional dialects of Bangla, such as Sylheti and Chittagonian, pose significant challenges for natural language processing due to their low-resource nature and substantial linguistic variation from standard Bangla. In this work, we present a systematic evaluation of eight open-source LLMs for translating fifteen distinct Bangla dialects into standard Bangla. To achieve this comprehensive coverage, we utilize a combination of established benchmarks and a novel dataset curated from an ongoing regional linguistic project. We assess model performance using a multi-metric framework that combines exact-match and error-rate evaluations such as, Averaged BLEU, WER, and CER with embedding-based semantic metrics including BERTScore, METEOR, and COMET. Additionally, we perform a detailed dialect-level linguistic analysis to identify the deep-seated structural, orthographic, and semantic barriers inherent to dialectal translation. Our study highlights the strengths and limitations of current open-source models, provides empirical insights for future dialect-aware fine-tuning, and contributes a reproducible benchmark for the research community.

Keywords: bangla dialects, low-resource language, translation, large language models

1. Introduction

Bangla exhibits immense geographic and linguistic diversity across a vast dialectal continuum, historically categorized into major groups like Rarhi, Varendri, Kamrupi, and Bangla (Chatterji, 1970). As illustrated in Figure 1[1], peripheral variants such as Sylheti, Chittagonian, and Rangpur differ significantly from standard Bangla (primarily Rarhi-based) in phonology, vocabulary, and morphology that they frequently border on mutual unintelligibility.

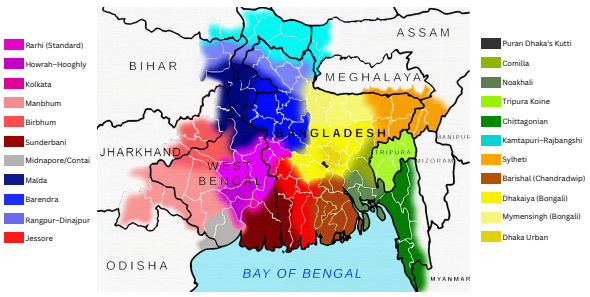

Figure 1: Regional dialects of the Bangla language.

This extreme variance creates a severe computational bottleneck. Table 5 in the appendix section provides examples that capture some of the regional dialectal diversity. IPA transcriptions are included solely for these instances to improve interpretability and are absent from the remaining data. Despite the rapid advancement of Large Language Models (LLMs), most foundation models are trained predominantly on standard Bangla, leaving these structurally different dialects largely underrepresented. Consequently, current Bangla NLP systems experience catastrophic performance drops often failing entirely at intent recognition, sentiment analysis, or semantic parsing when processing these peripheral variants. This failure carries a profound real world cost: millions of native dialect speakers are effectively excluded from modern digital inclusion initiatives, including voice assistants, e-governance platforms, and AI-driven healthcare, as their natural speech is systematically misinterpreted as erroneous. Translating these regional variants to standard Bangla is therefore an essential prerequisite for equitable downstream NLP performance.

However, this remains a fundamentally low-resource challenge characterized by a critical scarcity of aligned corpora. Furthermore, robust evaluation requires nuance. While the Averaged BiLingual Evaluation Understudy (BLEU) score serves as a widely adopted baseline for measuring n-gram translation precision, morphologically rich dialects demand evaluation methods that do not strictly penalize valid regional substitutions. The cost of relying exclusively on exact-match metrics

[1] Banerjee (2023), The Linguist Magazine

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 238–255
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

is substantial: it fundamentally misguides model development by artificially lowering the scores of models that successfully preserve semantic intent, thereby stalling progress in dialectal NLP. A comprehensive evaluation must therefore capture both exact-match surface normalization and deep semantic fidelity.

To address these limitations, we systematically evaluate open-source LLMs across fifteen distinct Bangla dialects. These specific dialects were strategically selected because they span the entire geographical and morphological continuum of Bangladesh capturing both mutually unintelligible linguistic islands (such as Chittagong and Sylhet) and central, transitional variants (such as Tangail and Mymensingh) thereby providing a holistic benchmark of true dialectal diversity. Leveraging a rigorously curated consolidation of existing public datasets, along with a private fold derived from an ongoing project referred to herein as the Regional-Standard Bangla Corpus (RSBC), we identify persistent structural translation barriers. Our evaluation employs a rigorous multi-metric approach. We combine surface-level lexical assessments such as Averaged BLEU (Papineni et al., 2002), Word Error Rate (WER), Character Error Rate (CER) with flexible and semantic-aware measures like Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Lavie and Agarwal, 2007), Crosslingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020), and BERTScore F1 (Zhang et al., 2020). Through detailed dialect-aware error analysis, this study provides a reproducible benchmark and critical insights, offering a roadmap for geographically inclusive, low-resource dialectal translation.

2. Related Work

The computational landscape for Bangla regional dialects is characterized by a significant disparity between data availability and task-specific utility. While dialects have been explored within socio-linguistic frameworks such as the *BIDWESH* (Fayaz et al., 2025) dataset introduced for multi-dialectal hate speech detection the development of parallel text corpora remains the most critical requirement for linguistic standardization and machine translation.

2.1. Evolution of Parallel Corpora and Translation Benchmarks

The primary obstacle to bridging dialects with standard Bangla is the lack of high-quality parallel data. Initial efforts to address this were often restricted to specific regions, such as *ChatgaiyyaAlap* (Chowdhury et al., 2025) for Chittagonian and *FeniVerse* (Mahi et al., 2025a) for the Feni dialect. Broader benchmarks like *Vashantor* (Faria et al., 2023) and the *BanglaDial* (Mahi et al., 2025b) corpus have since emerged to provide a multi-dialectal foundation. The *ONUBAD* (Sultana et al., 2025) dataset offers a comprehensive collection for automated conversion into standard Bangla, while recent works like *BhasaBodh* (Bhuiyan et al., 2025) provide parallel data for Chittagong and Sylhet, alongside addressing the real-world challenge of romanized digital communication. Despite these advancements, many sub-regional variations remain underrepresented, and the field continues to rely on a limited set of parallel pairs for model training.

2.2. From Traditional Neural Models to Large Language Models

Methodologically, dialect translation has shifted from rule-based mapping to Transformer-based architectures. Previous studies (Khandaker et al., 2025) have demonstrated that fine-tuned models like BanglaT5 can achieve a WER as low as 15.7% for standard-to-dialect translation.

Beyond text-based tasks, the complexity of Bangla dialects has also been investigated in the speech domain as well. For instance, Dipto et al. (2025) demonstrated that state-of-the-art ASR systems suffer severe performance drops on dialects like Chittagonian and Sylheti. Our benchmark mirrors this investigation in the text domain, evaluating whether LLMs exhibit the same vulnerability.

The current research frontier, however, centers on the deployment of LLMs. To adapt these foundation models, frameworks like *Sylheti-CAP* (Prama, 2025) have actively sought to improve zero-shot translation accuracy through sophisticated context-aware prompting. Systems like *BanglaDialecto* (Samin et al., 2024) have further attempted to unify speech recognition and translation into a single pipeline.

Building on this, recent work (Riad et al., 2025) has begun exploring parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) to develop dialect-aware conversational AI. For instance, Fine-tuned models such as Mistral, Gemma, and Claude for Question-Answering (QA) tasks across a subset of five regional variants. While these studies demonstrate the viability of fine-tuning LLMs for dialectal comprehension, they primarily rely on traditional n-gram metrics like BLEU and ROUGE for evaluation. As previously established, these exact-match metrics often fail to capture semantic fidelity in morphologically rich dialects. Furthermore, these efforts focus on chatbot-style QA generation rather than rigorous linguistic standardization. Consequently,

a comprehensive, multi-metric evaluation of open-source LLMs specifically targeting direct dialect-to-standard translation across a much wider array of geographically diverse Bangla dialects remains a critical gap in the literature.

3. Dataset

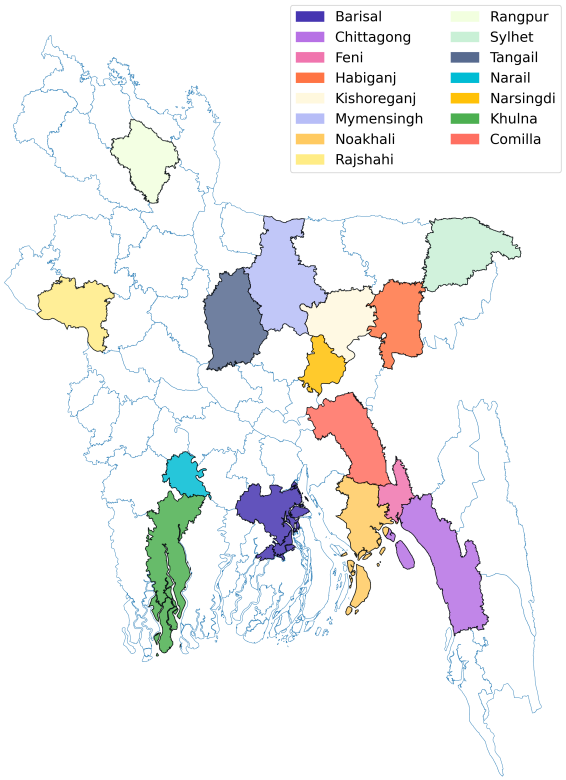

Figure 2: Map of evaluated dialects.

3.1. Task Definition

We formulate dialect-to-standard translation as a sequence-to-sequence generation task. A visual summary of the data conversion and preprocessing steps is presented in Figure 3. Given an input sequence of tokens in a regional Bangla dialect, the objective is to generate a semantically equivalent token sequence in Standard Bangla. The model must maximize the conditional probability of the standard translation given the dialectal input, ensuring deep semantic preservation while actively normalizing regional lexical, phonological, and morphological variations.

3.2. Proprietary Dialect Dataset

The RSBC corpus is composed of both aggregated open-source datasets and a private subset, which we designate as the Proprietary Dialect Dataset (PDD). This proprietary subset represents a curated checkpoint derived from a separate, ongoing research initiative dedicated to developing a large-scale Bangla parallel corpus. We integrated this private data into the current benchmark to augment our total sample size and expand our geographic coverage across a wider array of regions. While this specific subset remains temporarily proprietary pending the completion of its parent project, its inclusion ensures a more comprehensive, linguistically diverse, and statistically robust evaluation of the models. This PDD checkpoint currently encompasses data from 8 regions. To ensure high-quality parallel translations, we recruited 9 independent, native speakers for each respective regional dialect. The annotators were specifically trained for this task and spent approximately six months transcribing the data, while expert linguists continuously validated the accuracy of the outputs. All annotators were fairly compensated for their time.

3.3. Dataset Construction and Concatenation

To build a comprehensive multi-dialectal benchmark, we aggregated different publicly available sources such as *Vashantor* (Faria et al., 2023), *BIDWESH* (Fayaz et al., 2025), *Kothon* (Faisal et al., 2025), *BanglaDialecto* (Samin et al., 2024), *Bangla Regonal Text Corpus* (Ahmed et al., 2026), *Bangla Dialect Dataset* (Naeen et al., 2025), *FeniVerse* (Mahi et al., 2025a), *ChatgaiyyaAlap* (Chowdhury et al., 2025), *ONUBAD* (Sultana et al., 2025), in addition to the Proprietary Dialect Dataset (PDD) introduced in Section 3.2. In total, this benchmark encompasses 15 distinct regional Bangla dialects, including Chittagong, Noakhali, Barishal, Sylhet, Feni, Habiganj, Mymensingh, Khulna, Rangpur, Rajshahi, Tangail, Kishoreganj, Narail, Narsingdi, and Cumilla, as visualized in Figure 2. Each instance in the corpus provides a parallel sentence pair containing the original regional text, its Standard Bangla translation, and corresponding metadata tracking the region and dataset source. To ensure high data integrity across these disparate sources, all datasets were merged into a single unified corpus through a rigorous pipeline. This process included strict schema normalization, UTF-8 encoding standardization, a two-stage cross-dataset duplicate removal process utilizing both exact and fuzzy matching, and final manual and automated quality validation steps to remove malformed pairs.

The RSBC corpus contains 68,666 samples in total, with 17,004 derived from our proprietary dataset and 51,662 collected from external sources. Details regarding all these data sources

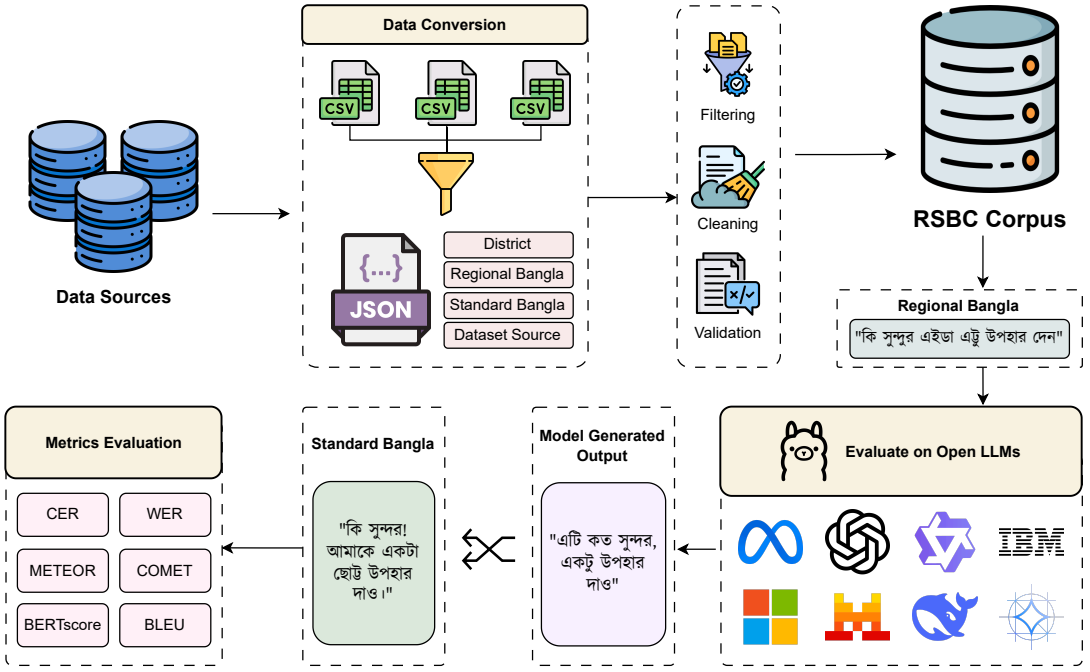

Figure 3: Overview of Regional–Standard Bangla Corpus (RSBC) creation and evaluation pipeline.

utilized in this work are presented in the Appendix (Section 12) at Table 3.

District	Samples	Total Words	Standard Unique	Dialect Unique	OOV (%)
Chittagong	19919	200524	19274	25225	23.59
Barishal	8065	80428	10386	11662	10.94
Noakhali	7983	89709	11492	13415	14.33
Rangpur	6108	40551	4677	7190	34.95
Habiganj	5357	50451	5564	8154	31.76
Sylhet	4904	36375	4305	6197	30.53
Feni	4035	42686	7865	8456	6.99
Kishoreganj	3623	21856	2844	4096	30.57
Mymensingh	2957	23467	3297	3927	16.04
Narsingdi	1973	14965	2416	3161	23.57
Narail	1143	6620	1467	1912	23.27
Khulna	909	4768	898	1266	29.07
Rajshahi	885	5624	1091	1621	32.70
Comilla	455	4145	768	1014	24.26
Tangail	350	2711	748	906	17.44
Total	**68,666**	**624,880**	**76,092**	**99,202**	**N/A**

Table 1: Dataset statistics and OOV% by district.

Table 1 summarizes the region-wise corpus statistics, including sample volume, vocabulary sizes, and the approximate Out-of-Vocabulary (OOV) percentage. Prior to analysis, texts were normalized and standard stopwords were removed.

Table 4 in the Appendix (Section 12) contains representative samples from our constructed RSBC corpus, illustrating regional dialect diversity and corresponding Standard Bangla translations. Figure 6 in the Appendix (Section 12) illustrates the distribution of samples across each regions.

As shown in Table 1, the aggregated corpus exhibits a severe class imbalance. High-resource dialects within our collection, such as Chittagong, contain nearly 20,000 samples, while the most low-resource variants, such as Tangail, contain only 350. A ratio of approximately 57:1. This disparity is an unavoidable reflection of the current scarcity of publicly available parallel corpora for peripheral Bangla dialects, necessitating reliance on the limited data currently available online and from our ongoing data collection efforts. Beyond sample distribution, two critical linguistic patterns emerge from this data.

First, for every evaluated region, the number of unique dialectal tokens significantly exceeds the unique tokens in the corresponding Standard Bangla translations (e.g., 25,225 dialect vs. 19,274 standard tokens in Chittagong). It is important to note that traditional Type-Token Ratio (TTR) is highly sensitive to corpus size, making cross-district comparisons of unique token counts (e.g., Chittagong vs. Tangail) susceptible to size-induced skew. However, because our corpus is strictly parallel, comparing the dialectal vocabulary directly against its Standard Bangla counterpart within the exact same set of sentence pairs perfectly controls for corpus size. This mathematically valid intra-dataset vocabulary inflation highlights the morphological richness and lack of standardized orthography in regional dialects, requiring models to learn complex, many-to-one mappings during translation.

Second, the OOV percentages reveal severe lexical bottlenecks. Regions such as Rangpur (34.95%), Rajshahi (32.70%), and Habiganj (31.76%) demonstrate that roughly one-third of their regional vocabulary is entirely disjoint from Standard Bangla.

This high prevalence of OOV tokens provides a direct, empirical explanation for why traditional

foundation models pre-trained almost exclusively on standard linguistic structures exhibit the elevated WER seen in Table 6 when attempting to normalize these peripheral variants.

To further quantify the structural and morphological divergence between the regional dialects and Standard Bangla, we analyzed the sentence length relationships across the parallel corpus. Figure 12 in the Appendix (Section 12) visualizes the character-level length distribution for all 15 evaluated dialects. In each subplot, a single data point represents a parallel sentence pair, plotted by its Standard Bangla length (x-axis) against its corresponding Regional Bangla length (y-axis). The solid diagonal line indicates absolute parity ($y = x$), where the dialect and standard sentences are identical in length.

While a strong positive correlation is naturally maintained across all districts, the visualizations reveal critical morphological variances. The dispersion of points away from the diagonal parity line illustrates that regional variations frequently utilize either highly compressed morphological constructs (falling below the line) or more verbose, descriptive phrasing (appearing above the line) to convey the same semantic meaning as the standard text.

Notably, highly divergent dialects such as Chittagong, Sylhet, and Noakhali exhibit broader scattering. Conversely, central and western dialects such as Tangail, Narail, Rajshahi, and Khulna demonstrate a much tighter clustering along the parity line. Specifically, the subplots for Rajshahi and Khulna reveal that their data points closely hug the diagonal, indicating that their sentence structures and morphological lengths strongly mirror Standard Bangla with minimal deviation. This variance mathematically demonstrates that translating peripheral Bangla dialects is not a symmetric, word-for-word substitution task. Instead, it inherently involves handling one-to-many and many-to-one token mappings. For foundation models (LLMs), this structural elasticity compounds the translation difficulty, as the model must bridge significant gaps in sentence length and syntactic structure to preserve the underlying intent, further reinforcing the necessity of semantic-aware evaluation metrics.

4. Dialect-Level Corpus Analysis

Beyond model evaluation, we conducted an exploratory linguistic analysis to quantify lexical and semantic variation across the fifteen regional dialects represented in our merged parallel corpus. Since this work is dialect-focused, understanding inter-dialect similarity is crucial for interpreting translation difficulty and model behavior.

4.1. Lexical Variation

To measure surface-level lexical overlap between dialects, we compute pairwise Jaccard similarity scores. For each region, we construct a vocabulary set consisting of unique word tokens extracted from the regional dialect side of the corpus.

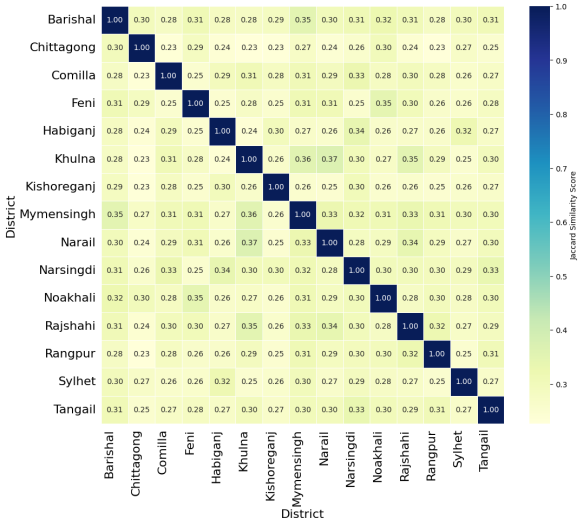

Figure 4: Jaccard Matrix of Bangla dialects.

The resulting Jaccard similarity matrix in Figure 4 reveals substantial lexical divergence across several geographically distant dialects, while neighboring regions tend to exhibit higher overlap. This supports the hypothesis that lexical variation contributes significantly to translation difficulty. While pairwise Jaccard similarity provides a useful heuristic for surface-level overlap, we note that the severe disparity in corpus sizes across districts (as outlined in Table 1) naturally suppresses the intersection ratio between high-resource and low-resource pairs. Consequently, these similarities represent corpus-level distributional overlap rather than absolute linguistic distance.

Specifically, the matrix demonstrates that Chittagong and Sylhet dialects exhibit the lowest similarity scores compared to Standard Bangla. These dialects function almost as linguistic islands, possessing distinct vocabularies and root words that severely limit surface-level overlap. In contrast, central and northern regions such as Tangail, Kishoreganj, and Rajshahi show markedly higher similarity indices, reflecting their closer morphological ties to the standard form.

4.2. Orthographic Divergence

To quantify the surface-level orthographic and morphological variation that models must overcome during normalization, we computed the average character-level Levenshtein distance between the sentences of each regional dialect and their corre-

sponding Standard Bangla translations. This metric measures the minimum number of character insertions, deletions, or substitutions required to transform one sentence into another, thereby capturing the spelling variations, affix differences, and phonological realizations inherent to each region.

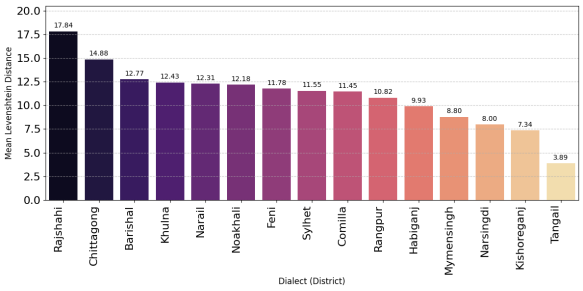

Figure 5: Average Character-Level Edit Distance to Standard Bangla.

The resulting bar chart in Figure 5 illustrates the mean edit distance to Standard Bangla across the evaluated regions. Higher values indicate a stronger orthographic divergence from the standard form. As shown, dialects from the southeastern and northeastern regions most notably Chittagong, Noakhali, Habiganj, and Sylhet exhibit the highest average edit distances. This reflects substantial phonological and morphological mutations that require heavy character-level alterations to standardize. In contrast, southwestern and central dialects such as Khulna, Narail, Cumilla, and Narsingdi demonstrate relatively lower edit distances, suggesting a much closer surface-form similarity.

Measuring the direct distance to the standard form provides a precise diagnostic of the normalization burden for each variant. This pronounced surface-level divergence offers a clear linguistic explanation for the elevated CER and WER observed in our zero-shot LLM evaluations, confirming that foundation models struggle heavily when confronted with severe morphological mutations, even if the underlying semantic intent is preserved.

4.3. Semantic Similarity Analysis

To evaluate meaning preservation and capture semantic closeness, we calculate the average cosine similarity between dialectal sentences and their Standard Bangla counterparts using multilingual Sentence-BERT embeddings (Reimers and Gurevych, 2019).

As reported in Table 2, similarity scores range from 0.632 in Chittagong to 0.897 in Tangail, revealing nuanced variations in semantic proximity. Regions such as Narail, Narsingdi, and Mymensingh also exhibit high alignment, showing scores of 0.822, 0.801, and 0.793, respectively. It suggests that they function as a linguistic "central hub" where core concepts and sentence structures remain highly preserved. In contrast, peripheral dialects like Chittagong, Sylhet, and Habiganj exhibit low semantic alignment, with scores falling to 0.632, 0.656, and 0.685, accordingly. This is particularly noteworthy because these regions are traditionally perceived as the most lexically distinct or even mutually unintelligible with Standard Bangla, a perception now validated by their positioning at the lower end of the semantic spectrum. These findings reinforce the critical research distinction between lexical divergence (surface-level word choice) and semantic divergence (the core meaning conveyed) when evaluating dialect-to-standard translation performance.

District	SSS	District	SSS
Tangail	0.896610	Rangpur	0.741638
Narail	0.821695	Kishoreganj	0.720103
Narsingdi	0.800928	Noakhali	0.716744
Mymensingh	0.793136	Rajshahi	0.688009
Khulna	0.782692	Comilla	0.686630
Feni	0.772380	Habiganj	0.685308
Barishal	0.743976	Sylhet	0.655511
Chittagong	0.632263		

Table 2: Average Semantic Similarity Scores (SSS) between regional dialects and standard Bangla.

Notably, these findings should be viewed through the lens of our dataset's sample distribution, which varies across regions. While the pronounced divergence in Chittagong and Sylhet highlights a significant linguistic gap, the high semantic proximity in Tangail suggests that these results serve as a high-signal indicator of how geographical proximity influences linguistic standardization. Ultimately, this observed semantic spread validates the hypothesis that translation difficulty is inherently dialect-dependent, necessitating more robust, dialect-aware modeling strategies to bridge these specific linguistic divides.

4.4. Lexical vs Semantic Correlation

To better understand how surface-level variation relates to deeper semantic alignment, we analyzed the correlation between lexical distance ($1 -$ Jaccard similarity) and sentence-level semantic similarity (Sentence-BERT). We observe a moderate negative correlation ($r = -0.47$) between lexical distance and semantic similarity across districts, indicating that increased lexical divergence generally leads to reduced semantic similarity. However, the relationship is not statistically significant ($p = 0.074$), suggesting that many dialectal variations preserve meaning despite substantial

lexical differences, see Appendix (Section 12.2) at Figure 7.

Rather than a weakness, this lack of strict statistical significance is a profound linguistic finding that perfectly supports our core hypothesis. As illustrated in Figure 7, while peripheral dialects (e.g: Chittagong or Sylhet) exhibit extreme vocabulary shifts, their semantic degradation is not perfectly linear. This proves that dialects are not simply "noisy" Standard Bangla, but robust, expressive linguistic systems capable of maintaining semantic integrity even when surface-level overlap is minimal. Consequently, this validates our argument that traditional exact-match metrics (which heavily penalize lexical divergence) are fundamentally inadequate for this task, and deep semantic-aware evaluations are absolutely necessary to judge true translation capabilities.

This corpus-level finding provides critical linguistic grounding for the observed LLM performance disparities: peripheral dialects do not merely suffer from surface-level spelling variations, but from deep-seated vocabulary shifts that fundamentally break the semantic mapping of standard foundation models. Consequently, this compounded divergence exponentially increases the translation burden, validating the severe spikes in WER observed in our evaluation. More importantly, it proves that standard zero-shot inference is fundamentally insufficient for bridging extreme dialectal gaps; resolving this semantic-lexical disconnect strictly necessitates the development of dedicated, dialect-aware fine-tuned LLMs capable of internalizing these robust regional linguistic structures.

5. Methodology

5.1. LLM Inference Framework

5.1.1. Problem Formulation

We are studying the task of translating regional Bangla text into standard Bangla across 15 distinct region dialects. Each problem instance consists of a regional Bangla input sequence, denoted as $R = \{r_1, r_2, \ldots, r_n\}$, where each r_i originates from a specific dialect within the region set $D = \{d_1, d_2, \ldots, d_{15}\}$. These inputs are paired with a fixed zero-shot instruction prompt, denoted as P_{zs}, which instructs the model to perform the dialect-to-standard conversion without relying on prior in-context examples. The objective is to generate the correct standard Bangla text, represented as $\hat{S}$, for each regional input r_i. This framework enables us to evaluate the inherent zero-shot translation capabilities of a LLM. Throughout the execution of our pipeline, the parameters for the model were kept entirely frozen. The entire procedure can mathematically be abbreviated as follows:

$$\hat{S}_i = LM([P_{zs}, r_i]) \quad (1)$$

Figure 3 illustrates a high-level overview of the proposed methodology.

5.1.2. Prompting Strategy

To ensure a robust and reproducible evaluation, we employ a structured zero-shot prompting framework (Kuo and Chen, 2023). Our methodology utilizes an instruction-based approach designed to enforce strict linguistic role conditioning and output constraints. Specifically, the prompt template incorporates structured tagging (e.g., `<translation>...</translation>`) to facilitate deterministic parsing and minimize both model hallucinations and extraneous conversational filler. To maintain experimental integrity and ensure a fair comparison across the 15 region-specific datasets, we apply an identical prompt configuration to all evaluated models. This standardized formatting ensures that performance variations are attributable to the models' inherent linguistic capabilities rather than prompt engineering biases. The comprehensive prompt template used throughout this study is detailed in Appendix 12.3.

5.2. Experimental Setup

5.2.1. Models

We evaluated a diverse suite of state-of-the-art LLMs to assess translation capabilities across various regional variants and dialects. Our selection encompasses both foundational and reasoning-optimized architectures, including `Llama3.2` (Grattafiori et al., 2024), `Phi4` (Abdin et al., 2024), `Qwen3` (Yang et al., 2025), `gpt-oss` (Agarwal et al., 2025), and `IBM Granite4` (Mishra et al., 2024). To further probe the limits of reasoning-heavy translation tasks, we included `DeepSeek-R1` (DeepSeek-AI, 2025), `Mistral` (Jiang et al., 2023), and `Gemma2` (Team et al., 2024). All models were deployed in a zero-shot prompting configuration to evaluate their inherent linguistic resilience and latent dialectal understanding without the influence of task-specific fine-tuning.

5.2.2. Evaluation Metrics

To rigorously quantify translation quality, we utilized a multi-metric framework designed to capture both lexical precision and semantic fidelity. For evaluating meaning preservation beyond surface-level overlap, we employed BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020), which

leverage contextual embeddings to ensure semantic alignment. We further utilized METEOR (Lavie and Agarwal, 2007) for its sensitivity to synonymy and stemming. To measure baseline n-gram translation precision, we report the averaged BLEU (Papineni et al., 2002) score, which captures structural mapping from unigrams up to 4-grams, while WER and CER were used to provide granular diagnostics of morphological and lexical normalization accuracy.

6. Result Analysis

6.1. Performance Analysis

The comprehensive evaluation detailed in Table 6 at Appendix (Section 12.4) reveals significant disparities in how current LLMs handle the structural and morphological complexities of regional Bangla. Specifically, the data highlights a severe gap between exact lexical normalization measured by BLEU, WER, and CER and the underlying semantic preservation captured by BERTScore and COMET.

`gpt-oss` and `Qwen3` emerge as the dominant performers across the macro-average. `gpt-oss` achieves the highest overall semantic alignment with a METEOR score of 0.3324, a COMET score of 0.8106 and a BERTScore of 0.8402, while also demonstrating the best exact-match BLEU average of 0.2224. `Qwen3` closely trails `gpt-oss` in semantic retention scoring a BERTScore of 0.8329, but secures the lowest overall CER of only 0.4333, indicating superior capability in handling fine-grained morphological standardizations. Both models showcase robust resilience when translating linguistically central dialects like Tangail, where `gpt-oss` reaches an impressive BERTScore of 0.8897 and a comparatively low WER of 0.6052.

However, it is crucial to interpret the macro-average reported in Table 6 at Appendix (Section 12.4) with caution due to the aforementioned dataset imbalance. The macro-average treats all fifteen districts equally, meaning the results from 350 samples of Tangail carry the same weight as those 19,919 samples from Chittagong. The exceptionally high performance figures for Tangail (e.g., `gpt-oss` achieving a BERTScore of 0.8897) likely reflect its closer linguistic proximity to Standard Bangla and its smaller, highly curated sample size, which is less representative of the true translation difficulty seen in heavily divergent dialects. Consequently, the unweighted macro-average may artificially inflate the perceived overall performance. While we report the macro-average to establish a baseline across all fifteen regions, future benchmarking on this dataset should consider weighted averages or isolated reporting of high-resource versus low-resource clusters to provide a fairer performance picture.

In contrast, `Mistral` and `Llama3.2` struggle profoundly with this dialectal normalization task. `Mistral` records the lowest macro-average across all positive metrics with a BLEU score of 0.0826, and COMET score of 0.5671. Additionally, it suffers from the highest global error rates and showed a WER of 1.2612, and CER of 1.0924. For heavily divergent regions such as Narsingdi and Kishoreganj, `Mistral`'s WER balloons to 1.6315 and 1.7807 respectively, suggesting massive token hallucination and structural breakdown.

To qualitatively investigate these extreme WER values, we analyzed Mistral's raw generation logs for these specific districts. We observed that instead of performing a direct mapping, Mistral frequently succumbed to severe output length mismatch driven by its failure to comprehend the peripheral dialect syntax. For example, given the Kishoreganj input "আন্নেরা দিয়া কারেন্ট আইছে?", the target Standard Bangla reference is "আপনাদের ঐদিকে কারেন্ট এসেছে?" (containing 4 tokens). However, Mistral completely failed the instruction and generated "আপনি একজট্রা হলো এক বিশ্ববর্তী ভাষা বিজ্ঞানী..." followed by a continuous, non-sensical loop of over 6,400 hallucinated tokens. This type of error manifesting as repetitive hallucination drastically inflates the insertion penalty component of the WER formula, mathematically pushing the score well above 1.0 and highlighting the brittleness of standard foundation models on low-resource regional variants.

`Llama3.2` fares only marginally better, indicating that without specialized fine-tuning, standard base models lack the necessary vocabulary and syntax mapping to bridge extreme regional variations. It is important to contextualize these results with respect to model size. The `gpt-oss (20B)` model possesses nearly seven times the parameter count of models like `Llama3.2 (3B)` and `Granite4 (3B)`. While `gpt-oss` demonstrates superior semantic retention, this is largely expected given its massive parameter advantage. The struggles of `Llama3.2` and `Granite4` highlight the limitations of highly compact models in low-resource dialectal mapping, rather than a fundamental architectural flaw.

Finally, the data reinforces our core linguistic hypothesis: exact-match metrics inherently penalize the necessary lexical substitutions required for dialect translation. Even the best performing model, `gpt-oss`, achieves a maximum average BLEU score of only 0.2224, yet its BERTScore average of 0.8402 proves that the fundamental meaning is

largely preserved.

Thus, while exact-match evaluations highlight the surface-level brittleness of LLMs, semantic-aware metrics are vital to accurately assessing true translational intent across the Bangla dialect continuum. The radar chart in Appendix (Section 12.2) at Figure 8 provides a holistic view of each model's multi-metric footprint. Because exact-match metrics inherently penalize dialectal normalizations, the "BLEU" axis remains universally contracted across all models. However, highly capable models like `gpt-oss` and `Qwen3` demonstrate a wide, expansive footprint toward the semantic axes, specifically BERTScore, COMET while successfully minimizing their stretch toward the error axes for instance, CER and WER, proving their superior capability in handling morphologically rich Bangla dialects.

Figure 9 and Figure 10 in Appendix (Section 12.2) separate the evaluation into positive semantic quality metrics including METEOR, COMET, BERTScore, BLEU and negative structural error metrics such as CER, WER. The quality distributions visually confirm `gpt-oss`'s dominance across all embedding-based metrics, closely followed by `Qwen3`. Conversely, the error distributions in Figure 10 highlight the extreme lexical brittleness of `Mistral` and `LLaMA 3.2`, which suffer from distinct spikes in both CER and WER, indicating frequent hallucinations and poor morphological mapping.

Figure 11 plots the models on a trade-off curve between semantic preservation (COMET) and surface-level structural degradation (WER). This scatter plot effectively clusters the models into three tiers of dialectal comprehension. `gpt-oss` and `Qwen3` occupy the optimal top-left quadrant (high semantic retention, low error), while `Mistral` is isolated in the bottom-right quadrant, representing a near-total failure to normalize the peripheral dialect structures effectively.

7. Future Work and Conclusion

While this framework successfully identifies critical linguistic barriers, current evaluations remain constrained by the severe geographical imbalance and representational disparity of available parallel data. Because we aggregated all currently available public data alongside our ongoing project, the resulting corpus suffers from heavy skew (e.g., Chittagong vs. Tangail). As an ongoing resource-building initiative, we are actively expanding our corpus to specifically target and over-sample these underrepresented sub-regional varieties to correct this imbalance. By using this expanded dataset, our objective is to fine-tune and publicly release a dedicated regional dialect-to-standard Bangla translation model, complete with comprehensive evaluation scripts and model cards. Ultimately, by providing both curated linguistic resources and an open-source baseline model, we aim to establish a sustainable infrastructure for dialect-aware NLP research and promote broader digital inclusion across all Bangla-speaking communities.

In this work, we presented a systematic evaluation of open-source LLMs across fifteen Bangla regional dialects, revealing significant semantic and structural challenges in dialect-to-standard translation. Through a multi-metric evaluation framework combining surface-level and semantic-aware, we provided a comprehensive analysis of current model capabilities and limitations in low-resource dialectal settings. By identifying these persistent translation gaps and proposing a unified benchmarking approach, this study contributes a foundational step toward more inclusive, linguistically grounded, and regionally representative language technologies.

8. Acknowledgement

The authors are deeply grateful to the annotators who have contributed to our past and present data collection initiatives. We extend a special thanks to those currently working on our private dataset; their contributions are vital to the ongoing expansion of regional Bangla dialect resources and have made this systematic evaluation possible. We also wish to thank the anonymous reviewers for their insightful and constructive feedback, which significantly strengthened the final version of this paper.

9. Ethical Statement

All datasets used in this study were obtained and handled in compliance with their respective licensing terms, with no personally identifiable information retained. The proprietary dataset was curated under appropriate oversight, with full acknowledgment of annotator contributions. Our work aims to support digital inclusion for underrepresented Bangla dialect communities, and we are committed to publicly releasing our benchmark and evaluation framework. The authors declare no conflicts of interest.

10. Bibliographical References

Marah I. Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R.

Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card.

Md. Tofael Ahmed Bhuiyan, Md. Abdur Rahman, and Abdul Kadar Muhammad Masum. 2025. BhasaBodh: Bridging Bangla dialects and Romanized forms through machine translation. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 113–118, Mumbai, India. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Tawsif Tashwar Dipto, Azmol Hossain, Rubayet Sabbir Faruque, Md. Rezuwan Hassan, Kanij Fatema, Tanmoy Shome, Ruwad Naswan, Md.Foriduzzaman Zihad, Mohaymen Ul Anam, Nazia Tasnim, Hasan Mahmud, Md Kamrul Hasan, Md. Mehedi Hasan Shawon, Farig Sadeque, and Tahsin Reasat. 2025. Are ASR foundation models generalized enough to capture features of regional dialects for low-resource languages? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 178–188, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and …. 2024. The llama 3 herd of models. *arXiv preprint*, 2407.21783.

Md. Rezuwan Hassan, Azmol Hossain, Kanij Fatema, Rubayet Sabbir Faruque, Tanmoy Shome, Ruwad Naswan, Trina Chakraborty, Md. Foriduzzaman Zihad, Tawsif Tashwar Dipto, Nazia Tasnim, Nazmuddoha Ansary, Md. Mehedi Hasan Shawon, Ahmed Imtiaz Humayun, Md. Golam Rabiul Alam, Farig Sadeque, and Asif Sushmit. 2025. Regspeech12: A regional corpus of bengali spontaneous speech across dialects.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Md. Arafat Alam Khandaker, Ziyan Shirin Raha, Bidyarthi Paul, and Tashreef Muhammad. 2025. Bridging dialects: Translating standard bangla to regional variants using neural models.

Hui-Chi Kuo and Yun-Nung Chen. 2023. Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 249–258, Toronto, Canada. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami Seelam, Brian Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David D. Cox, Ruchir Puri, and Rameswar Panda. 2024. Granite code models: A family of open foundation models for code intelligence. *arXiv preprint arXiv:2405.04324*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages

311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tabia Tanzin Prama. 2025. LLMs for low-resource dialect translation using context-aware prompting: A case study on Sylheti. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 292–308, Mumbai, India. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Md Jahid Alam Riad, Prosenjit Roy, Mahfuzur Rahman Shuvo, Nobanul Hasan, Stabak Das, Fateha Jannat Ayrin, Syeda Sadia Alam, Afsana Khan, Md Tanzim Reza, and Md Mizanur Rahman. 2025. Fine-tuning large language models for regional dialect comprehended question answering in bangla. In *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pages 1–6.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, and (and many others). 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, and Qwen3 technical report.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

11. Language Resource References

Ahmed, Md Tofael and Koli, Zannatul Mawa and Rahat, Azmain Mahtab and Akter, Taslima and Ayman, Umme. 2026. *BanglaRegionalTextCorpus: A curated dataset for four regional bangla dialects with standard Bangla and English translation*. PID https://doi.org/10.17632/92r62h4k5k.4.

Chatterji, Suniti Kumar. 1970. *The Origin and Development of the Bengali Language: Volume One*. Routledge, 1. PID https://doi.org/10.4324/9781003480945.

Chowdhury, Sinthia and Remal, Deawan Rakin Ahamed and Pasha, Syed Tangim and Islam, Ashraful and Noori, Sheak Rashed Haider. 2025. *ChatgaiyyaAlap: A dataset for conversion from Chittagonian dialect to standard Bangla*. Elsevier. PID https://doi.org/10.17632/wtms9xbkkw.1.

Faisal, Md. Atique and Sadaf, Farhan and Chowdhury, Dipta and Azrof, H.M. and Tanmay, Monojit Paul. 2025. *Kothon: A Large-Scale Dataset for Machine Translation of the Chittagonian and Sylheti Dialects into Standard Bangla*. Mendeley Data. PID https://doi.org/10.17632/2fv6vf9v2z.3. Version 3.

Faria, Fatema Tuj Johora and Moin, Mukaffi Bin and Wase, Ahmed Al and Ahmmed, Mehidi and Sani, Md. Rabius and Muhammad, Tashreef. 2023. *Vashantor: A Large-scale Multilingual Benchmark Dataset for Automated Translation of Bangla Regional Dialects to Bangla Language*. Mendeley Data. PID https://doi.org/10.17632/bj5jgk878b.2. ArXiv preprint arXiv:2311.11142.

Fayaz, Azizul Hakim and Uddin, MD. Shorif and Bhuiyan, Rayhan Uddin and Sultana, Zakia and Islam, Md. Samiul and Paul, Bidyarthi and Muhammad, Tashreef and Manzoor, Shahriar. 2025. *BIDWESH: A Bangla Regional Based Hate Speech Detection Dataset*. Mendeley Data. PID https://doi.org/10.17632/bpkrvf882k.1. ArXiv preprint arXiv:2507.16183.

Mahi, Mehraj Hossain and Khan, Anzir Rahman and Hoque, Zesanul and Mojumdar, Mayen Uddin. 2025a. *FeniVerse: A parallel corpus of Feni dialect, standard Bengali, and English*. Elsevier. PID https://doi.org/10.17632/25r923frfj.1.

Mahi, Mehraj Hossain and Khan, Anzir Rahman and Mojumdar, Mayen Uddin. 2025b. *BanglaDial: A merged and imbalanced text dataset for Bengali regional dialect analysis*. Elsevier. PID https://doi.org/10.17632/sm63ryv5dt.1.

Naeen, Md. Julkar and Das, Sourav Kumar and Supta, Supta Das Dip and Alahi, Tabib E

and Faisal, Md Faisal Tajwar and Rahat, Abdullah Al and Ayshe, Samurtha Jahan and Mahjabeen, Ummay and Tahmid, Md Tanvir and Mojumdar, Mayen Uddin. 2025. *Bangla Dialect Dataset: Exploring Linguistic Diversity Across Regions*. Mendeley Data. PID https://doi.org/10.17632/sm63ryv5dt.

Samin, Md. Nazmus Sadat and Ahad, Jawad Ibn and Medha, Tanjila Ahmed and Rahman, Fuad and Amin, Mohammad Ruhul and Mohammed, Nabeel and Rahman, Shafin. 2024. *BanglaDialecto: An End-to-End AI-Powered Regional Speech Standardization*. IEEE. PID https://doi.org/10.1109/BigData62323.2024.10826131.

Sultana, Nusrat and Yasmin, Rumana and Mallik, Bijon and Uddin, Mohammad Shorif. 2025. *ONUBAD: A comprehensive dataset for automated conversion of Bangla regional dialects into standard Bengali dialect*. Elsevier. PID https://doi.org/10.17632/6ft99kf89b.2.

12. Appendix

12.1. Dataset Sample Distribution

Dataset Source	Total Count	Covered Regions
Proprietary Dialect Dataset	17,004	Chittagong, Comilla, Habiganj, Kishoreganj, Narsingdi, Noakhali, Rangpur, Tangail
Vashantor	9892	Barishal, Chittagong, Mymensingh, Noakhali, Sylhet
BIDWESH	8985	Barishal, Chittagong, Noakhali
Kothon	7989	Chittagong
BanglaDialecto	4477	Noakhali
Bangla Regonal Text Corpus	4462	Barishal, Khulna, Narail, Rangpur
Bangla Dialect Dataset	4038	Barishal, Chittagong, Mymensingh, Rajshahi, Rangpur, Sylhet
Feniverse	4035	Feni
ChatgaiyyaAlap	4007	Chittagong
Onubad	2565	Barishal, Chittagong, Sylhet
Sylheti Dialects into Standard	1188	Sylhet
SylhetiCAP	24	Sylhet
Total	**68,666**	**12**

Table 3: Dataset-wise total sample count and covered districts (sorted in descending order).

Regions	Regional_bangla	Standard_bangla
Chittagong	আই হইদদে, হারে দইজজেদে?	আমি বলি যে কাকে ধরেছে?
Feni	হেতেরগো তুন অ্যাঁন্ডা ভালা খেইলছিলাম।	তাদের থেকে আমরা ভালো খেলেছিলাম।
Kishoreganj	ও আইচ্চা, তোরার বাইত গেলে কি খাইতারবাম?	ও আচ্ছা, তোদের বাড়িতে গেলে কি খেতে পারবো?
Sylhet	অউ ফুড়ি রীতিমতো মোর ভাই আর আম্মারে খষ্ট দের।	এই মেয়ে রীতিমতো আমার ভাই আর মাকে কষ্ট দিচ্ছে।
Mymensing	আমার কাল্কা একটা ভালা দাতের ডাক্তর দেহানির কথা।	আমার আগামীকাল একজন দাঁতের ডাক্তারের সাথে দেখা করার কথা।
Narsingdi	কিন্তু অয় যহন ঢুকছে দলে তহন থেইকা কিন্তু খুব ভালো খেলছে।	কিন্তু ও যখন ঢুকেছে দলে তখন থেকে খুব ভালো খেলছে।।
Barishal	হ্যায় যেমনে মোরে আশা করছে মুই নিজেরে হেমনেই সামনে আনছি।	সে যেভাবে আমাকে আশা করছে আমি নিজেকে সেভাবে উপস্থাপন করছি।
Khulna	কুথায় কুথায় ইংরেজি কবে আমাইগে সামনে হোডা আমাইগে সহ্য হবে না।	কখনও কখনও আমাদের ভাষার সাথে ইংরেজি মিশিয়ে ফেলা সহ্য করা হয় না।।
Rangpur	আজকা মুই তো হনু যে, আল্লাহর রাস্তায় যাইম, চল্লিশ দিন থাকিম, এই করিম।	আজকে আমি তো আল্লাহ্র রাস্তায় যাব, চল্লিশ দিন থাকব, এসবই করব।
Habiganj	তারফর ফাইয়াজের ফর ওইলো গিয়া আরেকটা আছে, ইটার নামটা বুইল্লা গেছিগা।	তারপর ফাইয়াজের পর হলো গিয়ে আরেকটা আছে এটার নাম ভুলে গিয়েছি।
Tangail	মধুপুরে রাবার বাগানের পরেই আছে মধুপুরের বিখ্যাত আনারসের হাট জলছত্র হাট।	মধুপুরে রাবার বাগানের পরেই আছে মধুপুরের বিখ্যাত আনারসের হাট, জলছত্র হাট।
Cumilla	এই কুত্তার বাচ্চা ধরতাম গিয়া আফনের অন্তত দুই থেকে তিন বার কুত্তার কামড় খাইছি।	এই কুকুরের বাচ্চা ধরতে গিয়ে আপনার অন্তত দুই থেকে তিন বার কুকুরের কামড় খেয়েছি।
Rajshashi	যখন সে হেবি একটা বেড নিউজ পেইয়ে গেলো তখন থেকেই সে খুব দুঃখ হোয়ে গেলো।	যখন সে খারাপ সংবাদ পেল, তখন তার দুঃখ হলো।
Noakhali	দুনিয়ার বুকে যেন দেওয়ান বাগির বিচার হয় যেন টাডা পরে আললা তোয়ার গজব নাজিল করে দাও।	দুনিয়ার বুকে যেন দেওয়ান বাগির বিচার হয় যেন টাডা পরে আললা তুমার গজব নাজিল করে দাও।
Narail	তুইতো ইবলিসকে আছিস রে ভাই! তোর খোঁজ করছিল ইবলিস। তোর নম্বরটা আই ইবলিসকে দি দিচি।	তুইতো ইবলিসকে আছিস রে ভাই! তোর খোঁজ করছিল ইবলিস। তোর নম্বরটা আমি ইবলিসকে দিয়েদিলাম

Table 4: Region-wise sample data from the RSBC corpus (sorted by Regional Bangla length)

Subset	Bengali text	IPA Transcription
Standard Bengali	সামনে ইদ, ইদের কেনাকাটা করতে হবে ভাই!	ʃɐmne ɪd̪, ɪd̪er kenɐkatɐ kor̪te hɔbe bʱɐ͡ɪ!
Rangpur	সামনোত ইদ, ইদের কেনাকাটি কইরবার নাগবে, ভাই!	ʃɐmnoʈo ɪd̪, ɪd̪er kenɐkatɪ ko͡ɪr̪bar nɐgbe, bʱɐ͡ɪ!
Kishoreganj	সামনে ঈদ, ঈদের কিনাকাডা করন লাগবো ভাই!	ʃɐmne ɪd̪, ɪd̪er kɪnɐkɐdɐ kɔron lɐgbo bʱɐ͡ɪ!
Narail	ছামনে ঈদ, ঈদির কিনাকাটা ওত্তি অবে ভাই!	cʰɐmne ɪd̪, ɪd̪ɪr kɪnɐkɐtɐ oʈʈ̪ɪ ɔbe bʱɐ͡ɪ!
Chittagong	সামনো দি ঈদ, ঈদর কিনাকাটা গরন ফরিবু বাই!	ʃɐmno d̪ɪ ɪd̪, ɪd̪or kɪnɐkɐtɐ gɔron pʰorɪbu bɐ͡i̯!
Narsingdi	সামনে ইদ, ইদের কিনাকাডা করুন লাগবো ভাই!	ʃɐmne ɪd̪, ɪd̪er kɪnɐkɐdɐ korun lɐgbo bʱɐ͡i̯!
Tangail	সামনে ইদ, ইদের কিনাকাটা করন নাগবো ভাই!	ʃɐmne ɪd̪, ɪdeɾ kenɐkɐtɐ kɔron nɐgbe bʱɐ͡ɪ
Habiganj	সামনে ইদ, ইদের লাইগ্গা কেনাখাডা খরোন লাগবো বাই!	ʃɐmne ɪd̪, ɪdeɾ lɐ͡ɪggɐ kenɐkʰadɐ kʰɔɾon lɐgbo bɐ͡i̯!
Sylhet	সামনে ইদ। ইদোর কেনাকাটা করা লাগবো, বাই।	ʃɐmne ɪd̪, ɪdoɾ kenɐkɐtɐ koɾɐ lɐgbo bɐ͡i̯.
Barishal	হোমনে ইদ, ইদের কেনাকাডা করতে অইবে বাই।	homne ɪd̪, ɪdeɾ kenɐkɐdɐ korte o͡ibe bɐ͡i̯.
Sandwip	সামনে ইদ, ইদের কিনাকাডা কইরতো হইব বাই।	ʃɐmne ɪd̪, ɪdeɾ kɪnɐkɐdɐ ko͡ɪr̪to ho͡ɪbo bɐ͡i̯.
Cumilla	সামনেদো ইদ, ইদের কিনাকাডা করতইবো বাই।	ʃɐmneɖo ɪd̪, ɪd̪er kɪnɐkɐdɐ korʈo͡ɪbe bɐ͡i̯.
Noakhali	সামনে ঈদ আইয়্যের, ঈদের বাজার হদাই করন লাইগবো ভাই।	ʃɐmne ɪd̪ a͡ɪer, ɪdeɾ baɟɐɾ hɔdoɪ kɔron lɐ͡ɪgbo bʱɐ͡i̯.

Table 5: Dialect diversification from different regions taken from (Hassan et al., 2025)

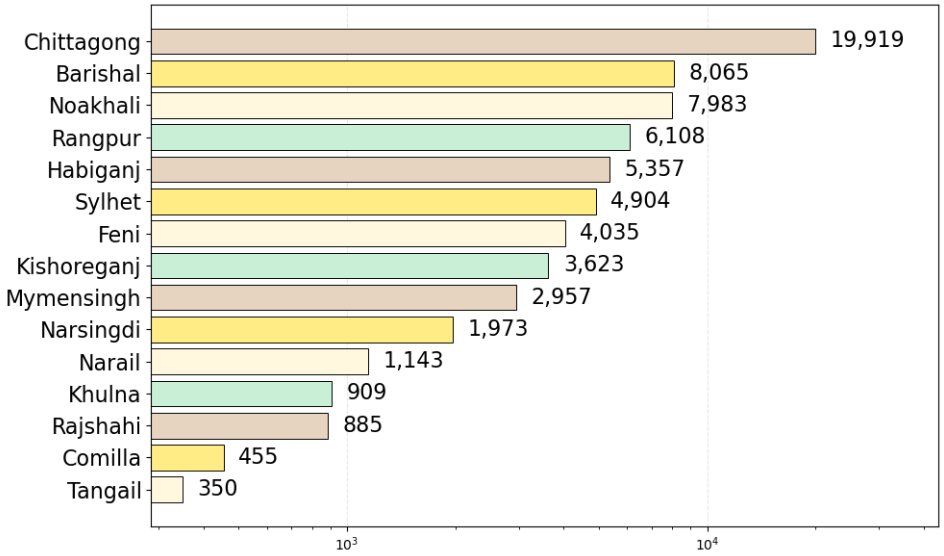

Figure 6: region-wise sample distribution.

12.2. Dialect-level Corpus Analysis

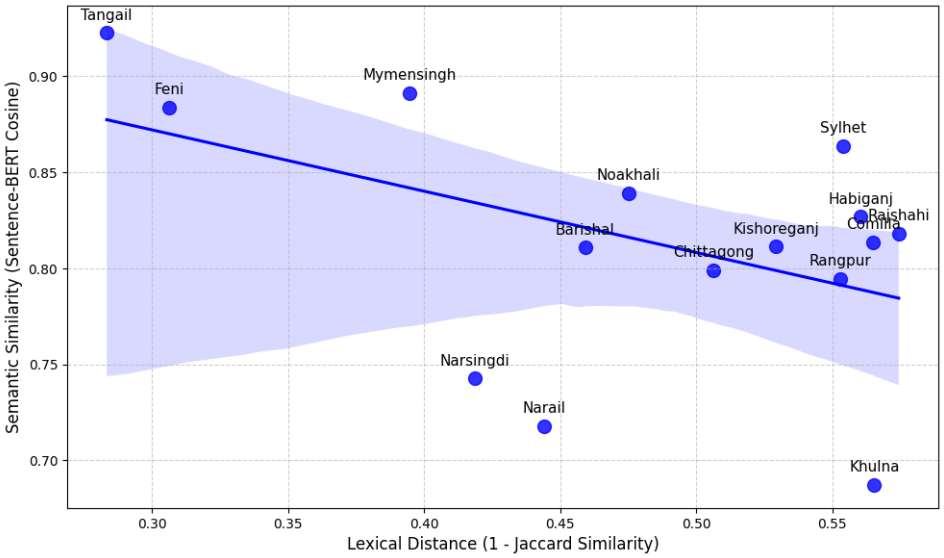

Figure 7: Correlation between Lexical Distance and Semantic Similarity across all the evaluated dialects

12.3. Prompt

The following prompt was utilized as the system instruction:

```
You are an expert linguist specializing in Bangladeshi regional
   dialects and Standard Bangla.

TASK: Translate the following {district} regional dialect text
into Standard Bangla (প্রমিত বাংলা).

CRITICAL INSTRUCTIONS:
- Provide ONLY the Standard Bangla translation.
- Do NOT provide any explanation, reasoning, or analysis.
- Do NOT include the original text in your response.
- Do NOT add any prefixes like "Translation:" or "Standard Bangla:".
- Output ONLY the translated text in Standard Bangla.
- Wrap your final translation in <translation>...</translation>
   tags.

REGIONAL DIALECT INPUT ({district}):
{regional_text}

STANDARD BANGLA TRANSLATION:
```

Here, `{district}` dynamically specifies the dialect source (e.g., Chittagong, Sylhet, Noakhali), and `{regional_text}` represents the dialectal input sentence.

12.4. Performance Analysis

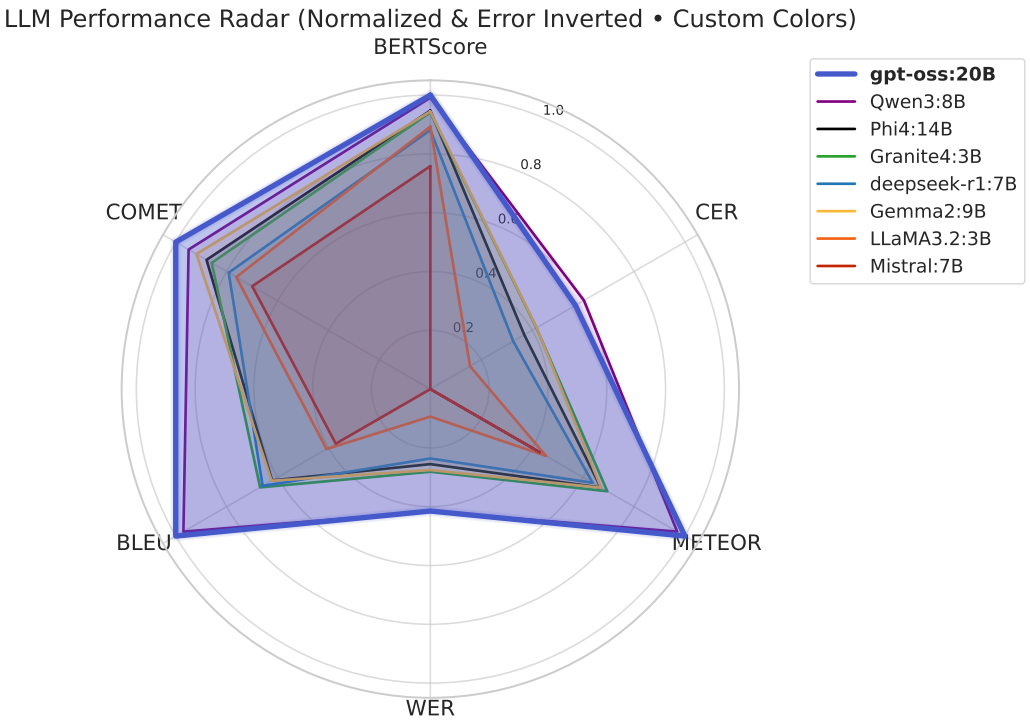

Figure 8: Multi-metric radar chart depicting the holistic performance footprint of each LLM.

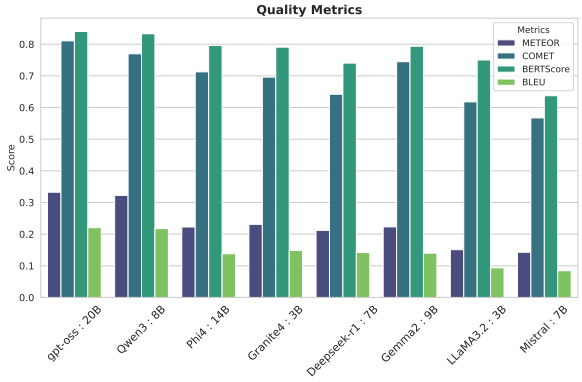

Figure 9: Macro-average distribution of quality metrics across evaluated LLMs.

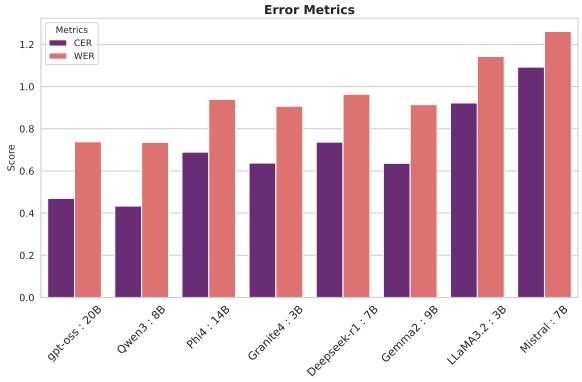

Figure 10: Macro-average distribution of error metrics across evaluated LLMs.

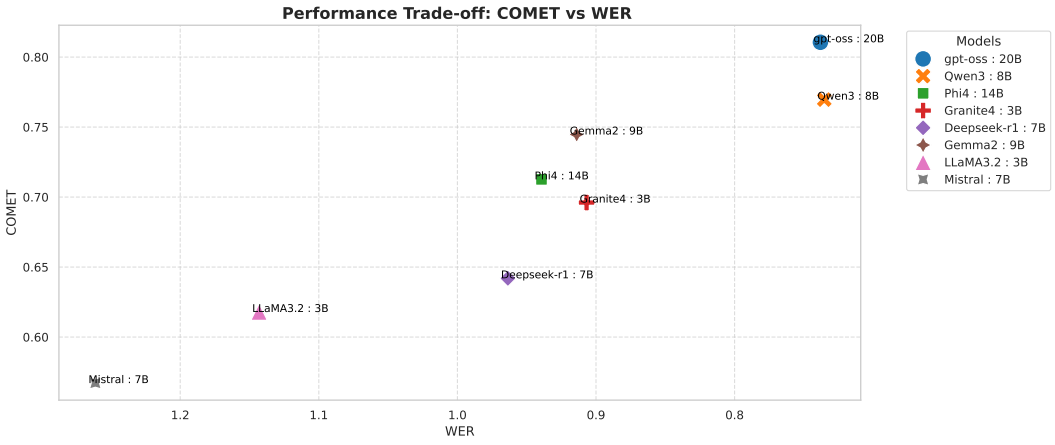

Figure 11: Performance Trade-off: Semantic alignment (COMET) versus morphological error (WER).

Models	Barishal						Chittagong					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.3893	0.8310	0.8515	0.2742	0.3957	0.6497	0.2712	0.7356	0.8245	0.1650	0.5537	0.8664
Qwen3:8B	0.3534	0.7782	0.8339	0.2471	0.4014	0.6786	0.2663	0.6796	0.8150	0.1797	0.4546	0.7728
Phi4:14B	0.2139	0.7067	0.7820	0.1366	0.6821	0.9313	0.1799	0.6422	0.7756	0.1111	0.7262	0.9764
Granite4:3B	0.2074	0.6283	0.7259	0.1416	0.6884	0.9123	0.1877	0.6145	0.7751	0.1206	0.6765	0.9519
deepseek-r1:7B	0.2399	0.6964	0.7933	0.1591	0.6125	0.8743	0.1586	0.5580	0.7257	0.1030	0.7682	1.0197
Gemma2:9B	0.2365	0.7431	0.7799	0.1556	0.6110	0.8787	0.1740	0.6697	0.7667	0.1099	0.7064	0.9763
LLaMA3.2:3B	0.1504	0.6046	0.7454	0.0961	0.9278	1.1340	0.1189	0.5532	0.7349	0.0724	0.9866	1.1912
Mistral:7B	0.1365	0.5514	0.6052	0.0848	1.0174	1.1803	0.1072	0.5095	0.5698	0.0640	1.1667	1.3058

Models	Cumilla						Feni					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2929	0.7511	0.8346	0.1787	0.4745	0.7699	0.3623	0.7983	0.8364	0.2409	0.3918	0.6807
Qwen3:8B	0.3043	0.7140	0.8415	0.1898	0.3932	0.7347	0.4261	0.7847	0.8504	0.3004	0.3483	0.6025
Phi4:14B	0.1933	0.6489	0.7863	0.1082	0.6855	0.9692	0.3156	0.7534	0.8182	0.2069	0.5041	0.7754
Granite4:3B	0.2237	0.6581	0.8009	0.1340	0.5786	0.8818	0.3055	0.7135	0.8094	0.2042	0.5164	0.7914
deepseek-r1:7B	0.2297	0.6310	0.7600	0.1451	0.7282	0.9617	0.3254	0.6573	0.7577	0.2296	0.5765	0.7716
Gemma2:9B	0.1884	0.6726	0.7775	0.1057	0.6767	0.9844	0.2625	0.7657	0.8010	0.1693	0.5546	0.8403
LLaMA3.2:3B	0.1308	0.5857	0.7485	0.0747	0.8689	1.1138	0.2490	0.6483	0.7772	0.1602	0.6993	0.9544
Mistral:7B	0.1410	0.5291	0.5858	0.0790	1.1389	1.2609	0.2022	0.5773	0.6593	0.1247	0.7252	0.9667

Models	Habiganj						Khulna					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2835	0.7482	0.8341	0.1624	0.5108	0.8163	0.3581	0.8921	0.8491	0.2423	0.4344	0.7071
Qwen3:8B	0.3163	0.7055	0.8381	0.1885	0.4074	0.7638	0.2718	0.8441	0.8083	0.1886	0.4981	0.8102
Phi4:14B	0.2020	0.6603	0.7892	0.1060	0.7139	0.9857	0.2578	0.7919	0.8274	0.1775	0.6122	0.8589
Granite4:3B	0.2317	0.6533	0.7966	0.1277	0.6288	0.9337	0.2013	0.7574	0.7877	0.1333	0.6601	0.9299
deepseek-r1:7B	0.2313	0.6083	0.7589	0.1359	0.7192	0.9853	0.1417	0.6832	0.7287	0.0986	0.7491	1.0051
Gemma2:9B	0.2038	0.6942	0.7896	0.1103	0.6734	0.9722	0.2763	0.8286	0.8290	0.1894	0.5412	0.8258
LLaMA3.2:3B	0.1529	0.5871	0.7460	0.0779	0.9211	1.1735	0.1160	0.6502	0.7346	0.0813	0.9189	1.1537
Mistral:7B	0.1553	0.5470	0.6174	0.0790	1.0835	1.2776	0.1255	0.6015	0.6820	0.0791	0.9470	1.1804

Models	Kishoreganj						Mymensingh					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2538	0.7509	0.8290	0.1353	0.5944	0.8570	0.4779	0.8827	0.8748	0.3535	0.3298	0.5498
Qwen3:8B	0.2746	0.7328	0.8392	0.1510	0.4746	0.8547	0.4425	0.8184	0.8630	0.3177	0.3520	0.5801
Phi4:14B	0.1661	0.6625	0.7894	0.0704	0.8746	1.1357	0.3011	0.7678	0.8070	0.2002	0.5395	0.7761
Granite4:3B	0.2039	0.6735	0.7977	0.1002	0.7104	1.0182	0.3146	0.7436	0.8210	0.2161	0.5150	0.7533
deepseek-r1:7B	0.1916	0.6267	0.7419	0.1039	0.9402	1.1339	0.2662	0.6399	0.7445	0.1853	0.6430	0.8386
Gemma2:9B	0.1757	0.6970	0.7883	0.0799	0.7253	1.0402	0.2935	0.8022	0.8099	0.1988	0.5241	0.7780
LLaMA3.2:3B	0.1141	0.5927	0.7295	0.0482	1.1160	1.3354	0.2350	0.6585	0.7861	0.1560	0.7177	0.9278
Mistral:7B	0.1328	0.5636	0.5971	0.0542	1.6930	1.7807	0.1871	0.5769	0.6810	0.1188	0.8730	1.0497

Models	Narail						Narsingdi					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.3822	0.9075	0.8583	0.2665	0.4030	0.6580	0.2456	0.7967	0.8270	0.1592	0.4869	0.8326
Qwen3:8B	0.3055	0.8684	0.8235	0.2141	0.4542	0.7494	0.2575	0.7746	0.8058	0.1869	0.4145	0.7858
Phi4:14B	0.2941	0.8198	0.8410	0.1994	0.5760	0.8179	0.1538	0.6770	0.7816	0.0960	0.7752	1.0661
Granite4:3B	0.2453	0.7960	0.7982	0.1632	0.6368	0.9042	0.1679	0.6935	0.7703	0.1138	0.6857	0.9929
deepseek-r1:7B	0.1878	0.7238	0.7407	0.1311	0.7198	0.9787	0.1856	0.6521	0.7181	0.1341	0.7861	1.0208
Gemma2:9B	0.3077	0.8530	0.8393	0.2069	0.5309	0.7920	0.1585	0.7113	0.7800	0.1007	0.6698	0.9841
LLaMA3.2:3B	0.1472	0.6932	0.7558	0.1016	0.8309	1.0796	0.1111	0.6118	0.7277	0.0741	1.1189	1.3586
Mistral:7B	0.1213	0.6035	0.6349	0.0785	0.9595	1.1718	0.1167	0.5870	0.6432	0.0759	1.3655	1.6315

Models	Noakhali						Rangpur					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2728	0.7464	0.8060	0.1740	0.4841	0.7921	0.3375	0.8279	0.8441	0.2141	0.4610	0.7521
Qwen3:8B	0.3240	0.7299	0.8274	0.2214	0.3661	0.6837	0.3033	0.7741	0.8314	0.1932	0.4699	0.8031
Phi4:14B	0.1836	0.6576	0.7664	0.1135	0.6575	0.9271	0.2185	0.7258	0.8033	0.1256	0.7375	0.9971
Granite4:3B	0.2140	0.6488	0.7799	0.1383	0.5916	0.8730	0.2302	0.7159	0.8006	0.1387	0.6454	0.9459
deepseek-r1:7B	0.2083	0.6013	0.6999	0.1392	0.6520	0.8842	0.1819	0.6327	0.7315	0.1124	0.7879	1.0360
Gemma2:9B	0.1853	0.6887	0.7599	0.1165	0.6463	0.9324	0.2243	0.7567	0.8023	0.1324	0.6567	0.9502
LLaMA3.2:3B	0.1513	0.5780	0.7453	0.0949	0.8810	1.0899	0.1198	0.6109	0.7382	0.0664	1.0389	1.2561
Mistral:7B	0.1396	0.5323	0.6062	0.0854	0.8956	1.0965	0.1359	0.5812	0.6472	0.0727	1.1609	1.3468

Models	Rajshahi						Sylhet					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.2894	0.8385	0.8117	0.1938	0.5615	0.7809	0.3107	0.7970	0.8324	0.2180	0.5306	0.7533
Qwen3:8B	0.2702	0.7810	0.8063	0.1843	0.5832	0.8208	0.2785	0.7310	0.8240	0.1908	0.4738	0.7593
Phi4:14B	0.1961	0.7406	0.7728	0.1273	0.7174	0.9381	0.1914	0.6862	0.7748	0.1232	0.7865	1.0028
Granite4:3B	0.1878	0.7134	0.7752	0.1254	0.7201	0.9499	0.1957	0.6641	0.7861	0.1266	0.7233	0.9630
deepseek-r1:7B	0.1400	0.6180	0.7080	0.0926	0.9205	1.0803	0.1419	0.5772	0.7052	0.0953	0.8471	1.0445
Gemma2:9B	0.2034	0.7892	0.7762	0.1337	0.6670	0.9136	0.2126	0.7337	0.7840	0.1418	0.6464	0.8958
LLaMA3.2:3B	0.1379	0.6345	0.7505	0.0894	0.8842	1.0913	0.1311	0.5912	0.7513	0.0848	1.0082	1.2087
Mistral:7B	0.1147	0.5787	0.6716	0.0737	1.0092	1.1852	0.1208	0.5516	0.6648	0.0753	1.0815	1.2430

Models	Tangail						Average					
	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓	METEOR ↑	COMET ↑	BERTScore ↑	BLEU ↑	CER ↓	WER ↓
gpt-oss:20B	0.4594	0.8546	0.8897	0.3282	0.4359	0.6052	**0.3324**	**0.8106**	**0.8402**	**0.2224**	0.4699	0.7381
Qwen3:8B	0.4409	0.8282	0.8852	0.3093	0.4079	0.6300	0.3223	0.7696	0.8329	0.2157	**0.4333**	**0.7353**
Phi4:14B	0.2714	0.7489	0.8234	0.1700	0.7497	0.9322	0.2226	0.7126	0.7959	0.1382	0.6892	0.9393
Granite4:3B	0.3482	0.7673	0.8349	0.2473	0.5808	0.8005	0.2310	0.6961	0.7906	0.1487	0.6372	0.9068
deepseek-r1:7B	0.3488	0.7237	0.7907	0.2574	0.5995	0.8189	0.2119	0.6420	0.7403	0.1465	0.7367	0.9636
Gemma2:9B	0.2407	0.7648	0.8195	0.1496	0.7121	0.9449	0.2229	0.7447	0.7935	0.1387	0.6361	0.9139
LLaMA3.2:3B	0.1977	0.6687	0.7806	0.1267	0.9129	1.0782	0.1509	0.6179	0.7501	0.0907	0.9221	1.1431
Mistral:7B	0.2059	0.6162	0.6930	0.1223	1.2685	1.2414	0.1428	0.5671	0.6372	0.0826	1.0924	1.2612

Table 6: District-wise and macro-average performance of LLMs for dialect-to-standard Bangla translation across fifteen districts. The best average metrics are in bold.

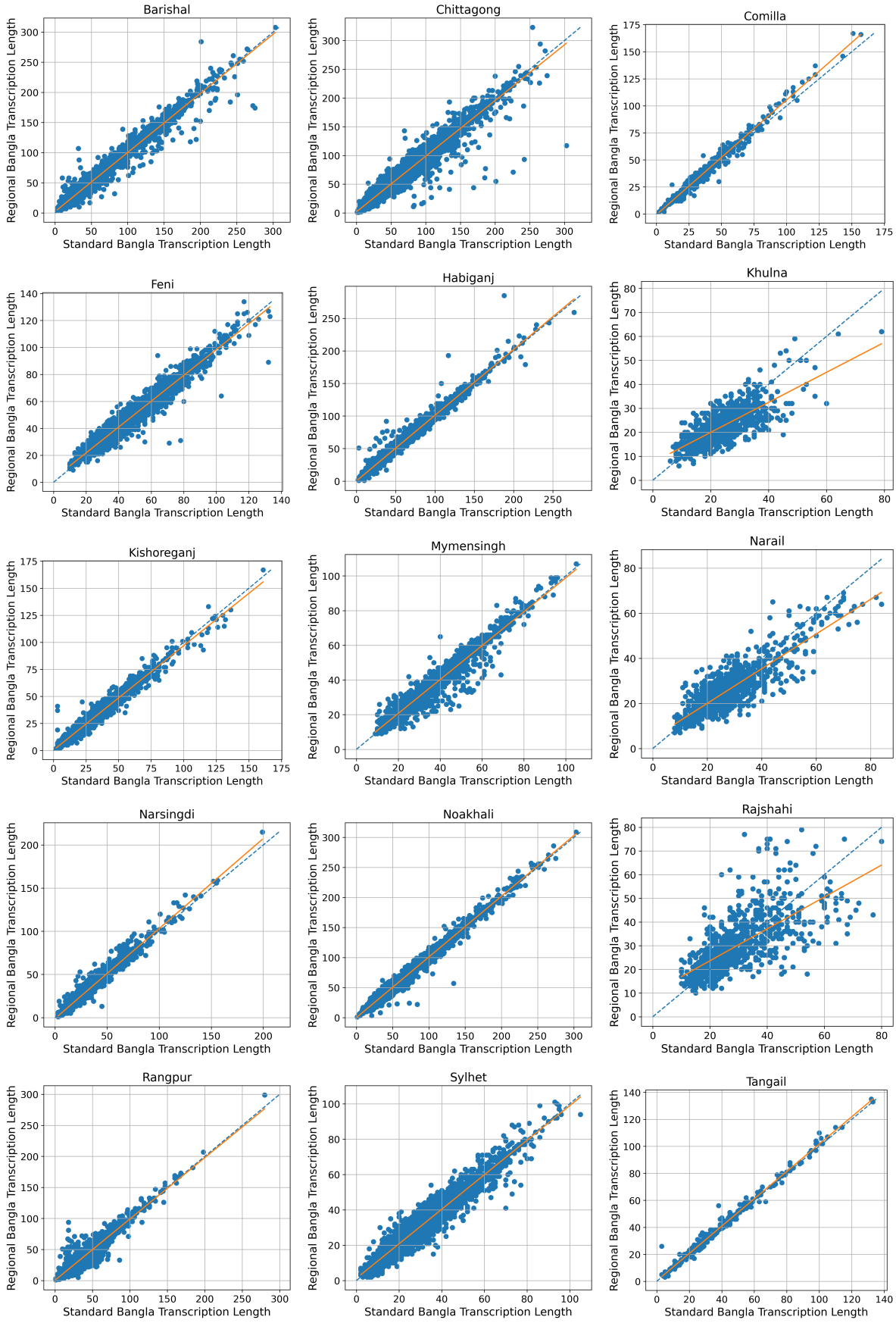

Figure 12: Sentence length comparison between Regional and Standard Bangla across dialects.