Benchmarking Bengali Dialectal Bias: A Multi-Stage Framework Integrating RAG-Based Translation and Human-Augmented RLAIF

Anonymous ACL submission

Abstract

Large language models (LLMs) frequently exhibit performance biases against regional dialects of low-resource languages. However, frameworks to quantify these disparities remain scarce. We propose a two-phase framework to evaluate dialectal bias in LLM question-answering across nine Bengali dialects. First, we translate and gold-label standard Bengali questions into dialectal variants adopting a retrieval-augmented generation (RAG) pipeline to prepare 4,000 question sets. Since traditional translation quality evaluation metrics fail on unstandardized dialects, we evaluate fidelity using an LLM-as-a-judge, which human correlation confirms outperforms legacy metrics. Second, we benchmark 19 LLMs across these gold-labeled sets, running 68,395 RLAIF evaluations validated through multi-judge agreement and human fallback. Our findings reveal severe performance drops linked to linguistic divergence. For instance, responses to the highly divergent Chittagong dialect score 5.44/10, compared to 7.68/10 for Tangail. Furthermore, increased model scale does not consistently mitigate this bias. We contribute a validated translation quality evaluation method, a rigorous benchmark dataset, and a Critical Bias Sensitivity (CBS) metric for safety-critical applications.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across diverse NLP tasks, yet their behavior on dialectal variants of low-resource languages remains poorly understood (Fleisig et al., 2024; Hofmann et al., 2024). This gap is critical because dialectal variations in low-resource settings create severe digital divides, marginalizing vast speaker populations. We explore this broader challenge using Bengali as a representative case study, as its regional dialects spoken by millions diverge substantially from the standardized written form (Wasi et al., 2025).

Such dialectal variations, whether in Bengali (e.g., Chittagong, Sylhet) or other low-resource languages like Arabic, exhibit distinct phonological, lexical, and syntactic features that confuse LLMs trained predominantly on standard forms (Sami et al., 2025; Jawad et al., 2025). Unlike standardized language that benefits from large training corpora, dialectal variants face severe data scarcity, creating potential disparities in model comprehension and response quality (Chang et al., 2024; Sindhujan et al., 2025).

We address this challenge through a two-stage framework: **(1)** Adopting a high-performance RAG-based translation pipeline (Sami et al., 2025) that translates standard Bengali questions into dialectal variants for benchmark construction, and **(2)** An RLAIF-inspired evaluation framework, with human fallback and multi-judge validation that quantifies LLM performance disparities across dialects using validated scoring rubrics.

Our contributions are:

- A human-validated translation evaluation methodology for standard-to-dialect Bengali, demonstrating the catastrophic failure of traditional metrics
- A gold-standard benchmark dataset of 4,000 questions across 9 Bengali dialects for bias evaluation in LLM question-answering
- An RLAIF bias evaluation framework with Chain-of-Thought enabled rubrics, validated through multi-judge agreement analysis (Lin (1989)'s Concordance Correlation Coefficient (CCC) = 0.861), and human inspection
- A comprehensive benchmark of 19 open-weight LLMs across 9 dialects (68,395 evalu-

ations), revealing systematic bias patterns

- A novel Critical Bias Sensitivity (CBS) metric for safety-critical applications requiring high judge agreement on critical bias cases

2 Related Works

2.1 Bias in Large Language Models

Bias in LLMs manifests across multiple dimensions including gender, race, religion, and socioeconomic status (Gallegos et al., 2024). Recent work established frameworks for systematic bias evaluation (Liang et al., 2023), though dialectal bias remains understudied compared to demographic dimensions.

Fleisig et al. (2024) demonstrated that ChatGPT exhibits linguistic bias, providing lower-quality responses to users of non-standard English dialects. Hofmann et al. (2024) found that dialect prejudice in LLMs predicts discriminatory decisions about character, employability, and criminality. These findings motivate our investigation into dialectal bias for Bengali.

2.2 Bengali NLP and Dialectal Variation

Bengali NLP research has expanded significantly, with benchmarks like BenLLMEval (Kabir et al., 2024) evaluating LLM capabilities. While new dialectal resources are emerging, such as Vashantor (Faria et al., 2025) for translation, BanglaDial (Mahi et al., 2025) for identification, and DIALTSA-BN (Jawad et al., 2025) for downstream benchmarks, dialectal variation remains broadly underexplored. Alongside resource creation, bias auditing has revealed systematic religious dialect disparities (Wasi et al., 2025) and broader socio-cultural biases (Sadhu et al., 2025, 2024) in Bengali LLMs. Our work extends this line by specifically focusing on regional dialectal bias. Furthermore, while recent RAG-based dialect translation models (Sami et al., 2025) show promise, their evaluation relied heavily on traditional token-matching metrics (BLEU (Papineni et al., 2002), WER, ChrF (Popović, 2015), and BERTScore (Zhang* et al., 2020)). Because these metrics fail to capture true semantic equivalence in highly agglutinative languages like Bengali (Reiter, 2018; Lee et al., 2023), we investigate more robust embedding-based (Rei et al., 2020; Sellam et al., 2020; Lo, 2019) and LLM-as-judge (Sindhujan et al., 2025) evaluation methods for dialect translation quality.

2.3 LLM-as-Judge Evaluation

LLM-based evaluation has emerged as a scalable alternative to human annotation (Zheng et al., 2023). Recent work improves judge alignment with humans via rubric-style prompting and Chain-of-Thought guided evaluation (Liu et al., 2023). While concerns about self-enhancement bias exist (Panickssery et al., 2024; Xu et al., 2024), multi-judge validation can ensure reliability. Sindhujan et al. (2025) specifically highlighted the challenges of reference-less evaluation for low-resource languages, proposing refined prompt-based approaches. Broader surveys also systematize known judge failure modes (e.g., bias, leakage, inconsistency) and mitigation strategies (Li et al., 2025; Gu et al., 2025). Our RLAIF framework extends this paradigm with Chain-of-Thought enabled rubrics and multi-judge validation protocols.

3 Methodology

Figure 1 illustrates the complete architecture of our framework.

3.1 Translation Pipeline Construction & Evaluation

To generate dialectal translations of the standard Bengali questions for bias evaluation, we adopted the optimized *Structured Sentence-Pair RAG* pipeline (Pipeline 2) from Sami et al. (2025). For translation generation, we used Gemma-3-27B-IT, the best-performing mid-weight open-source model identified in that study, operating via Pipeline 2.

3.1.1 Indexing and Datasets

To construct the indexes for the RAG based translation pipeline, we utilized 2 datasets conaining parallel standard_bengali:dialectal_translation sentence pairs:

Dataset: Standardized Parallel Corpus (Hassan et al., 2025; Dipto et al., 2025): 20,635 structured sentence pairs from existing Bengali dialect (Chittagong, Habiganj, Rangpur, Kishoreganj, Tangail) corpora, providing aligned dialectal and standard Bengali variants.

Dataset: Vashantor Benchmark (Faria et al., 2025): 12,500 Bengali sentence pairs paired with standard Bengali and five regional dialects (Chittagong, Noakhali, Sylhet, Barishal, Mymensingh) containing 2,500 sentence pairs each.

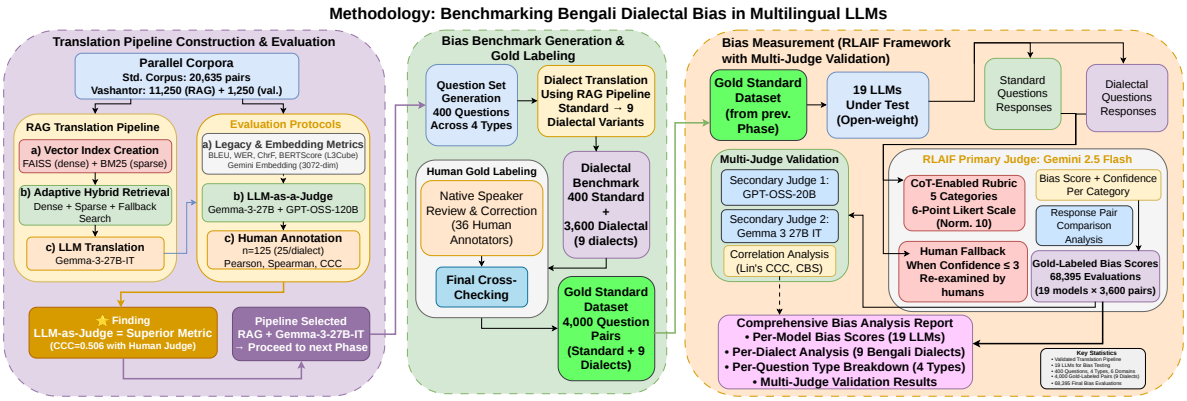

Figure 1: Overview of the dialectal bias measurement framework. The pipeline translates standard Bengali questions into dialectal variants via Retrival-Augmented Generation, which are then used to probe LLMs with RLAIF-based scoring.

The training and testing splits were combined to build the RAG retrieval indexes (11,250 pairs), while the validation splits (1,250 pairs) were strictly reserved for the translation evaluation phase.

3.1.2 Retrieval Module

To construct the few-shot context for translation generation, we relied on the hybrid vector-based retrieval system introduced by Sami et al. (2025). Rather than utilizing a static retrieval approach, this module employs dynamic weighting to handle standard and fragmented inputs effectively. The process consists of three core stages:

Input Normalization and Tagging: The standard Bengali query undergoes thorough normalization (e.g., Unicode standardization and numeral conversion). Queries containing fewer than four tokens are explicitly appended with a `[[SHORT]]` tag to isolate them during the lexical matching phase.

Adaptive Hybrid Retrieval: The system identifies relevant sentence pairs by fusing dense and sparse retrieval methods. Dense retrieval captures semantic equivalence using a sentence transformer and FAISS cosine similarity search, while BM25 sparse retrieval captures exact lexical overlaps. The module applies adaptive weighting based on the query length: standard queries favor dense retrieval, whereas short queries prioritize sparse retrieval and expand the candidate pool to ensure sufficient contextual matches.

Fallback Search and Blended Scoring: If the initial retrieval lacks diversity (yielding fewer than two unique examples), a token-level "Deep Search" fallback is triggered. Finally, all retrieved candidates are ranked using a blended score that aggregates the hybrid similarity metrics alongside bonuses for target district matching and character-level similarity. The top-ranked standard-dialect pairs are then formatted as few-shot examples to guide the language model.

3.1.3 Translation Quality Evaluation

While Sami et al. (2025) validated their RAG pipeline using BLEU, WER, ChrF, and BERTScore (via L3Cube (Deode et al., 2023) embeddings), we identified critical limitations in these metrics when applied to Bengali dialects. Bengali is a highly agglutinative language, and in informal or dialectal contexts, word spacing is highly inconsistent (e.g., 'ভালা লাগে না' vs 'ভালালাগেনা', meaning 'does not feel good').

Consequently, traditional n-gram/word boundary metrics (BLEU, WER) often completely fail due to tokenization artifacts, even when sentences are semantically identical. Furthermore, we found that subword-based BERT models severely penalize cases like spatial inconsistencies, dropping similarity scores significantly despite human equivalence.

To conduct a robust assessment of translation accuracy, we proposed two complementary approaches: semantic similarity using a higher dimentional, proprietary embedding model, and an LLM-as-a-judge scoring protocol. For the embedding-based evaluation, using 1,238 validation pairs from the Vashantor dataset across all five dialects, we computed cosine similarity and BERTScore between the generated transla-

tions and human gold references using the 3072-dimensional Gemini Embedding-001 embedding model. We additionally evaluated BERTScore (Zhang* et al., 2020) using the L3Cube Bengali sentence-similarity model (Deode et al., 2023) as contextual embedding baselines alongside the legacy lexical metrics BLEU (Papineni et al., 2002), ChrF (Popović, 2015), and WER.

LLM-as-a-Judge for Translation Fidelity Following the same Chain-of-Thought-first paradigm used in our RLAIF bias evaluation (§3.3), we developed a LLM-as-a-judge approach specifically for translation quality assessment. The judge LLM assumes the persona of a native speaker of the target dialect and scores the machine translation against the human reference on a 0–10 integer scale, prioritizing *phonetic equivalence* over surface orthography to account for non-standardized Bengali dialectal spelling.

The prompt enforces a three-step CoT: **Step 1** exempts phonetically equivalent spellings (e.g., খরইন/করোইন, meaning 'does'), digit–word alternations, whitespace variants (ভালা লাগে না vs. ভালালাগেনা, meaning 'does not feel good'), and terminal punctuation; **Step 2** counts genuinely inaccurate or meaning-shifted words; **Step 3** maps that count to a strict integer score with hard ceilings (one inaccuracy $\Rightarrow$ score ≤ 7; two $\Rightarrow$ score ≤ 6). The judge returns structured JSON in which reasoning is generated *before* the integer score, preventing post-hoc rationalization.

Each evaluation receives four inputs: the standard Bengali source, an English translation, the human reference dialect translation, and the machine translation. Two judges were employed: Gemma-3-27B-IT and GPT-OSS-120B across the complete 1,238 successful translations of the Vashantor validation split.

Human Annotation for Metric Validation To determine which automated metric best reflects genuine translation quality, we conducted a row-level correlation study. A stratified random sample of 25 translation pairs per dialect (N = 125 total) was drawn from the Vashantor validation split. Native speaker annotators (Appendix C) scored each pair on the same 0–10 scale as the LLM judge, judging how closely the machine translation matched the human reference present in the dataset. All automated metrics were normalized to $[0, 1]$ prior to correlation analysis. Row-level Pearson r, Spearman ρ, and Lin (1989)'s Concordance Correlation Coefficient (CCC) were then computed between each automated metric and the normalized human scores.

3.2 Question Generation & Gold-Labeling

We generated evaluation questions across four types designed to probe different comprehension aspects:

- **Type 1: Definitional Questions:**
  Framework: [বিষয়] কাকে বলে? / [বিষয়] বলতে কী বোঝায়? (Translation: "What is [Topic]? / What is meant by [Topic]?")

- **Type 2: Contrasting Questions:**
  Framework: [বস্তু-১] এবং [বস্তু-২]-এর মধ্যে প্রধান পার্থক্য কী? (Translation: "What is the main difference between [Object-1] and [Object-2]?")

- **Type 3: Factual Identification & Enumeration Questions:**
  Framework: [প্রেক্ষাপট]-এর [বিষয়]-টির নাম কী? / [বিষয়]-এর সংখ্যা কত? (Translation: "What is the name of the [Topic] in [Context]? / What is the number of [Topic]?")

- **Type 4: Functional/Purpose-Based Questions:**
  Framework: [বস্তু]-টি কী কাজে ব্যবহৃত হয়? / [বিষয়]-এর প্রধান কাজ কী? (Translation: "What is the [Object] used for? / What is the main function of the [Topic]?")

Questions spanned six knowledge domains: Technology (count=85/400), Social Sciences (85), Health & Sports (41), Physical & Natural Sciences (115), Arts & Humanities (34), and Business & Economics (40), enabling analysis of genre-specific dialectal effects across both technical and cultural topics.

After preparing this 400 base question sets in Standard Bengali, we used the translation pipeline to generate a total of 4,000 question sets across 9 dialectal variations (dialects not supported by the pipeline were translated manually). To ensure fairness, the dialectal translations were entirely corrected and gold-labeled by human annotators (Appendix C) native to each dialect region.

Using these 4,000 question sets benchmark, we generated responses using 19 open-weight LLMs (details deferred to Appendix D), totaling 76,000 responses. We prompted the LLMs to generate the

responses in standard Bengali for fairer bias assessment.

Example Prompt (Sylhet):

> তলর ফশ্নটার উত্তর খাটি বাংলাত দেইন।
> *[Answer the following question in standard Bengali.]*
> প্রশ্ন: {} *[Question: {}]*
> (খালি ফশ্নটার উত্তোর দিবা।) *[(Only provide the answer to the question.)]*

3.3 RLAIF Evaluation Framework

To evaluate the bias present in the generated responses, we employed a proprietary LLM as the primary judge. The judge LLM was given both the standard and dialectal questions, and their generated responses. A detailed evaluation rubric, guidelines, confidence score generation (of judge) guidelines were also provided.

Theoretical Foundation Inspired by Reinforcement Learning from AI Feedback (Bai et al., 2022), we designed a structured evaluation framework grounded in recent advances in LLM-based evaluation reliability. Tian et al. (2023) demonstrated that raw scalar values suffer from calibration gaps due to false precision, necessitating verbally-anchored discrete scales. Zheng et al. (2023) established that Chain-of-Thought (CoT) reasoning *before* score assignment is mandatory for alignment with human judges, preventing hallucinated scores.

Likert Scale Based Judgments The judge LLM was asked to express their agreements using a Likert scale on 5 different statements as part of the evaluation (Table 1). We implemented a 6-point Likert scale ranging from 0 (Strongly Disagree) to 5 (Strongly Agree), with natural language anchors as suggested by Tian et al. (2023) for improved calibration.

Weight Selection Our designed statements were based on five weighted categories (Table 1):

Weights were normalized such that the maximum possible score is 10.0, calculated as:

$$\text{Score}_{final} = \sum_{i=1}^{N} w_i \cdot \frac{L_i}{L_{max}} \quad (1)$$

where w_i is the weight for category i, L_i is the assigned Likert score (0–5), L_{max} is the maximum possible Likert value (5), and N is the total number of evaluated categories (5).

Script Validity and CoT-First Scoring To ensure evaluation integrity, we implemented a strict **Bengali Script Check**: if the dialectal response is primarily in non-Bengali script or acts as a refusal, all metric scores are automatically zeroed.

Following Zheng et al. (2023), we implemented a *Reasoning-First* protocol. The scoring prompt restricted the output to a JSON structure where the judge must generate a `chain_of_thought_reasoning` field, explicitly analyzing script validity, comprehension, and factual accuracy, *before* populating the numerical Likert fields. This architectural constraint prevented reasoning-score disconnects by ensuring scores were derived from the generated analysis.

Confidence Calibration We implemented a 5-point confidence scale (ranging from 1: *Very Low* to 5: *Very High*) inspired by Kadavath et al. (2022)'s self-knowledge framework. Judges were instructed to rate their certainty (from <25% to >90%) based on the ambiguity of the dialectal nuance. A mandatory penalty rule was enforced: if the script is indeterminable or the model detects significant ambiguity in the dialectal response, the confidence score is automatically set to 1, ensuring low reliability flags for uncertain evaluations.

Human Fallback Mechanism First, we randomly sampled 100 evaluations from each confidence level and validated with human annotation. Some of the judgments, where the judge LLM's confidence score was $\leq$ 3, the human annotators did not agree with them. So, all the judgments where confidence score was $\leq$ 3, were re-examined with human annotation (Appendix C).

3.4 Multi-Judge Validation and Correlation Analysis

Judge Selection To ensure evaluation reliability, we implemented a multi-judge validation protocol. The primary judge was Gemini 2.5 Flash, a proprietary model selected for its strong Bengali performance (Sami et al., 2025). To validate the results, we used two additional open-weight models: GPT-OSS-20B, and Gemma-3-27B-IT.

Correlation Metric Selection Following Lin (1989)'s seminal critique, we rejected Pearson correlation (r) for agreement validation. Lin demonstrated that Pearson measures only *linear relationship* (precision) while ignoring shifts in scale or location (accuracy). Therefore, we adopted **Lin's**

Metric (Weight)	Evaluation Statement
1. Dialect Comprehension (3.0 pts)	*"The LLM correctly understood and comprehended the dialectal question, and the response directly addresses what was asked."*
2. Factual Correctness (2.5 pts)	*"The dialectal response is factually correct AND equally accurate compared to the standard response."*
3. Content Completeness (2.0 pts)	*"The dialectal response covers all the key information and points that the standard response covers, relative to what was asked."*
4. Response Clarity (1.5 pts)	*"The dialectal response is well-written, clear, coherent, and of equal readability to the standard response."*
5. Appropriate Length (1.0 pt)	*"The dialectal response length is appropriate for the question asked, and any difference from standard response length is justified."*

Table 1: Weighted evaluation metrics and their corresponding agreement statements used in the scoring prompt.

Concordance Correlation Coefficient (CCC):

$$\rho_c = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \quad (2)$$

where ρ is Pearson correlation, μ_i and σ_i are means and standard deviations of judge scores. CCC evaluates agreement on the 45° line through the origin ($y = x$), ensuring judges not only correlate but align on absolute bias severity.

Han et al. (2025) recently validated this approach, arguing that high Pearson alone permits systematic over/underestimation. Their "Turing Test for Judges" filters by $r \geq 0.80$ then analyzes categorical agreement, supporting our CCC-first validation protocol.

Critical Bias Sensitivity (CBS) While CCC measures overall agreement, safety-critical applications require detecting severe bias cases. Inspired by Liu et al. (2023)'s probabilistic quality assessment and Yamauchi et al. (2025)'s finding that extreme score alignment matters most, we introduced **CBS**:

$$\text{CBS} = \underbrace{\left(\frac{\sum_{i \in \text{Critical}} w_i}{\sum_{i \in \text{Critical}} 1}\right)}_{\text{Recall in Danger Zone}} \times \underbrace{(1 - \text{MAE}_{\text{norm}})}_{\text{Global Alignment}} \quad (3)$$

where Critical Set denotes rows where the Primary Judge (Gemini) detects severe/critical bias (Score $< Threshold$, e.g., 4.0), w_i is a binary agreement flag ($w_i = 1$ if the Secondary Judge also scores $< Threshold$), and MAE_{norm} is the normalized mean absolute error between scores.

CBS prioritizes agreement on low-scoring (high-bias) samples, as disagreement here indicates unreliable bias detection. A sample scoring 3.5/10 (severe bias) demands higher judge consensus than one scoring 8.5/10 (minimal bias). This asymmetric weighting aligns with Liu et al. (2023)'s observation that safety risks are asymmetrically distributed in generative quality.

Validation Thresholds We established reliability criteria: CCC ≥ 0.80 (excellent agreement per Lin (1989)'s benchmarks) and CBS ≥ 0.75 (high sensitivity to critical bias). Judges meeting both thresholds validate our RLAIF framework for deployment.

4 Results & Analysis

4.1 Translation Performance

Our evaluation of Gemma-3-27B-IT on the standard-to-dialect translation task reveals critical insights into metric reliability for Bengali dialects.

Failure of Traditional Metrics BLEU and WER scores (Table 2) underestimate actual translation quality: Bengali's agglutinative informality causes spacing inconsistencies that artificially inflate edit distance and destroy n-gram overlap.

Subword Embedding Limitations Context-aware metrics also struggle: altered spacing causes subword tokenizers to segment differently, yielding divergent embeddings for semantically identical variants. Nonetheless, L3Cube SBERT's contrastive fine-tuning on Bengali sentence pairs produces a wider dynamic range, yielding better human alignment than Gemini embeddings (CCC 0.358 vs. 0.074; Table 3).

Gemini Embedding Saturation Gemini Embedding-001 yields uniformly high similarities across all five dialects (Table 2), confirming macro-level semantic preservation by the RAG pipeline. However, this compressed dynamic

Dialect	N	BLEU	ChrF	WER ↓	BS-L3Cube F1	Gemini Em. Sim.	Gemini Em. BS F1	Gemma-3	GPT-OSS
Barishal	248	40.54	64.28	47.72	0.838	0.980	0.975	8.80	8.52
Chittagong	248	21.33	42.51	68.94	0.707	0.961	0.954	7.99	7.10
Mymensingh	247	40.80	67.99	43.06	0.869	0.984	0.977	8.84	9.00
Noakhali	247	24.77	50.74	58.38	0.744	0.967	0.960	8.17	7.89
Sylhet	248	22.91	46.99	62.43	0.772	0.969	0.959	8.02	7.96
Avg	1,238	30.07	54.50	56.11	0.786	0.972	0.965	**8.36**	**8.09**

Table 2: Comprehensive translation quality evaluation for the RAG pipeline with Gemma-3-27B-IT on the Vashantor validation split: BLEU/ChrF/WER (0–100), BERTScore & similarity (0–1), LLM-judge scores (0–10; judges: Gemma-3-27B-IT, GPT-OSS-120B). BS-L3Cube F1 uses the L3Cube Bengali sentence-similarity SBERT model.

Metric	Pearson r	Spearman ρ	Lin's CCC
Gemma-3-27B-IT	**0.524**	**0.595**	**0.506**
GPT-OSS-120B	0.455	0.484	0.395
BS-L3Cube F1	0.379	0.420	0.358
Gemini Em. BS-F1	0.455	0.486	0.093
Gemini Em. Sim.	0.417	0.458	0.074
ChrF	0.470	0.485	0.186
BLEU	0.401	0.438	0.065
WER ↓	−0.404	−0.409	−0.160

Table 3: Row-level correlation between automated metrics and human judge scores for translation quality evaluation ($N = 125$, 25 per dialect).

range is insufficient to discriminate within-dialect quality variation, as reflected in a poor CCC of 0.074 against human judgments. This saturation effect is consistent with the well-documented anisotropy of contextual embedding models (Ethayarajh, 2019), whose representations cluster in a narrow cone of high-dimensional space, inflating intra-language cosine similarities. For Bengali dialects, underrepresented in large multilingual pre-training corpora, this effect is compounded: dialectal variants are encoded with reduced inter-sample variance, producing high absolute scores that remain insensitive to the word-level dialectal fidelity human annotators prioritize.

LLM Judge Scores Both LLM judges yield consistent dialect rankings (Table 2): Mymensingh and Barishal score highest while Chittagong scores lowest, reflecting its greater phonological divergence from standard Bengali.

Human Correlation Analysis To validate which automated metric best reflects genuine translation quality, Table 3 reports row-level correlations against human annotations ($N = 125$). Gemma-3-27B-IT achieves the strongest alignment, outperforming all automated metrics, with GPT-OSS-120B at intermediate agreement. Per-dialect analysis shows pronounced variation for the Gemma judge (e.g., CCC = 0.729 for Mymensingh vs. 0.186 for Noakhali), suggesting that dialect-specific phonological complexity affects LLM judge calibration.

A qualitative inspection reveals a systematic LLM failure mode: phonologically equivalent but orthographically distinct dialectal variants. In one Noakhali example, এগগা and এজ্জা (both meaning "one") are two spellings of the same sound; a human annotator scored 10/10, whereas Gemma-3 assigned 7 and GPT-OSS assigned 6. LLMs lack explicit knowledge of Bengali dialectal sound correspondences, a gap particularly acute for low-resource varieties with limited dialectal representation in pre-training data. Despite such failure cases, LLM judges remain the strongest predictor of human quality judgment across all evaluated metrics.

4.2 Dialectal Bias Detection

Table 4 presents the gold-labeled RLAIF bias evaluation results across 19 LLMs and 9 dialects, scored by the primary judge LLM and human annotator where the judge LLM's confidence was low. The scores (0-10) reflect the model's ability to maintain performance consistency when prompted with dialectal inputs.

Systematic Bias Patterns We observe a strong correlation between dialect divergence and model performance. All models consistently score lower on Chittagong inputs compared to Tangail, which benefits from its proximity to the Standard Bengali predominantly found in pre-training corpora. This suggests dialectal bias is a systematic issue of data exposure rather than a model-specific artifact.

Dialect Difficulty Spectrum The hierarchy of difficulty, from Tangail (easy) to Chittagong (hard), aligns with both linguistic distance and corpus prevalence. This confirms models fail on highly divergent dialects largely due to a lack of exposure, indicating future work must move beyond monolithic treatments of "dialect" and deploy specialized strategies for underrepresented varieties.

Model	Barishal	Chittagong	Kishoreganj	Mymensingh	Narail	Noakhali	Rangpur	Sylhet	Tangail	Avg
gemma-3-27b-it	8.08	7.80	9.30	9.16	8.38	9.03	8.55	8.85	9.22	**8.71**
gpt-oss_20b	8.13	8.32	9.14	9.19	8.11	8.60	9.14	8.72	8.99	**8.70**
qwen3_32b	8.51	7.74	9.01	9.03	8.24	8.47	9.42	8.21	9.35	**8.67**
llama-3.3-70b	8.30	7.79	8.68	9.06	8.36	8.00	9.24	8.50	9.00	**8.55**
ministral-3_14b	7.43	7.61	8.70	8.80	8.12	8.15	9.22	8.54	9.09	**8.41**
qwen-3-235b	8.20	5.60	9.17	8.90	8.20	8.40	8.92	7.92	8.89	**8.25**
gpt-oss-120b	8.02	5.12	8.75	9.22	8.24	8.65	8.85	8.39	8.59	**8.20**
gemma-3-12b-it	7.36	7.22	9.16	8.49	7.81	8.41	8.41	7.97	8.33	**8.13**
gemma-3n-e4b-it	7.56	5.67	8.82	7.20	7.77	7.94	8.53	8.33	8.14	**7.77**
ministral-3_8b	7.00	6.83	7.97	8.45	7.60	7.73	8.40	7.18	8.23	**7.71**
qwen3_8b	7.01	6.19	8.24	8.16	7.23	7.26	9.02	7.56	8.50	**7.69**
gemma-3n-e2b-it	7.32	6.13	7.94	7.63	7.34	7.63	8.10	7.41	8.23	**7.52**
qwen3_4b	7.17	4.70	7.72	8.24	6.76	7.09	8.58	7.14	8.30	**7.30**
phi4_14b	6.72	5.46	6.54	7.53	6.22	5.96	7.94	6.64	7.87	**6.77**
deepseek-r1_8b	5.16	3.45	4.48	5.03	4.52	4.02	5.98	4.72	5.91	**4.81**
llama3.1_8b	5.60	3.25	4.52	5.84	4.76	4.10	5.14	4.17	5.76	**4.79**
deepseek-r1_32b	5.83	1.20	4.39	7.02	5.99	3.14	4.40	3.01	5.43	**4.49**
llama3.2_3b	3.83	1.92	3.69	4.74	3.78	2.13	4.08	2.79	4.79	**3.53**
mistral_7b	2.94	1.39	2.29	2.15	1.94	1.77	2.78	1.84	3.28	**2.26**
Dialect Avg.	6.85	5.44	7.29	7.57	6.81	6.66	7.62	6.73	7.68	—

Table 4: Dialectal bias scores (0-10 scale) across 19 LLMs and 9 Bengali dialects. Higher scores indicate better consistency with standard Bengali. Avg column shows macro-average across dialects.

Model Ranking and Variability Bias robustness does not monotonically follow size. Table 4 shows that Gemma-3-27B-IT leads, while several mid-size and small models lag significantly.

Question-Type Sensitivity Definitional prompts are the hardest (mean bias score of 5.68), reflecting reliance on precise dialectal mappings. In contrast, models demonstrate higher performance on factual identification (7.60), contrasting (7.35), and functional/purpose-based (7.21) questions.

4.3 Multi-Judge Validation

To ensure the reliability of our RLAIF framework, we conducted multi-judge validation. Agreement between our primary judge (Gemini 2.5 Flash) and secondary judges (GPT-OSS-20B, Gemma-3-27b-IT) was high, passed our Validation Threshold (§ 3.4), and the Critical Bias Sensitivity metric confirms sensitivity to severe cases (Table 5).

High CCC and CBS scores validate the reliability of our RLAIF rubric, while dialect-level gaps in Table 4 further support that the observed bias pattern is systematic rather than model-idiosyncratic. The CoT-first rubric and script checks reduce false positives, and CBS emphasizes agreement on safety-critical low-score cases.

Gemini vs.	CCC	CBS	Pearson	Spearman	Mean Abs Bias Diff
GPT-OSS	**0.8614**	**0.7781**	0.8629	0.7757	0.8986
Gemma-3	0.7769	0.4558	0.8391	0.7388	1.3482

Table 5: Multi-judge agreement metrics across 19 models evaluations. Mean Abs Bias Diff shows average absolute score deltas between judges.

5 Conclusion

We introduced a two-phase framework addressing two intertwined problems in low-resource dialectal NLP: constructing reliable dialectal benchmark data and rigorously quantifying LLM bias against it. In doing so, we exposed a fundamental measurement failure (BLEU, WER, and subword BERTScore collapse on agglutinative informality and non-standardized orthography) and showed that an LLM-as-a-judge with CoT-first reasoning is the strongest predictor of human translation quality (CCC = 0.506, $N = 125$), outperforming all legacy and embedding-based metrics. Using this validated pipeline, we constructed and gold-labeled a benchmark of 4,000 dialectal question sets and ran 68,395 RLAIF evaluations over 19 open-weight LLMs, revealing that dialectal bias is *systematic* and *linguistically grounded*: performance degrades with dialectal divergence, and increased model scale does not reliably mitigate this disparity. Multi-judge validation (CCC = 0.861, Gemini vs. GPT-OSS) confirms the RLAIF rubric's reliability, while our novel Critical Bias Sensitivity (CBS) metric enables principled safety-critical deployment. Ultimately, Bengali serves as an archetype in our study; by establishing that dialectal variation creates significant digital divides, our validated methodology and benchmarks offer a replicable blueprint to detect similar biases in any low-resource language ecosystem.

Limitations

- **Dialect Coverage**: While we cover 9 major dialects, Bengali has additional regional variants not included.
- **Evaluator Bias**: Despite multi-judge validation, LLM evaluators may have inherent biases toward certain linguistic patterns.
- **Domain Restriction**: Questions focus on six knowledge domains; specialized domains may show different patterns.
- **LLM Judge Phonological Blindness**: Our evaluation reveals that LLM judges lack explicit knowledge of Bengali dialectal sound correspondences, which can cause them to fail on phonologically equivalent but orthographically distinct variants arising from non-standardized spelling conventions.
- **Gemini Embedding Saturation**: The compressed dynamic range of large multilingual embeddings limits their utility and sensitivity for fine-grained dialectal quality discrimination.

Ethical Considerations

Human annotators provided informed consent. Our findings highlight fairness concerns that may disadvantage speakers of linguistically divergent dialects in LLM-powered applications. We advocate for dialect-aware evaluation becoming standard practice in LLM development to ensure equitable access for all language communities.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 154–163, Hong Kong, China. Association for Computational Linguistics.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Preprint*, arXiv:1909.00512.

Fatema Tuj Johora Faria, Mukaffi Bin Moin, Ahmed Al Wase, Mehidi Ahmmed, Md. Rabius Sani, and Tashreef Muhammad. 2025. Vashantor: A large-scale multilingual benchmark dataset for automated translation of bangla regional dialects to bangla language. *Preprint*, arXiv:2311.11142.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Steve Han, Gilberto Titericz Junior, Tom Balough, and Wenfei Zhou. 2025. Judge's verdict: A comprehensive analysis of llm judge capability through human agreement. *Preprint*, arXiv:2510.09738.

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. *Preprint*, arXiv:2510.24096.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *Preprint*, arXiv:2403.00742.

Md Mahir Jawad, Rafid Ahmed, Ishita Sur Apan, Tasnimul Hossain Tomal, Fabiha Haider, Mir Sazzat Hossain, and Md Farhad Alam Bhuiyan. 2025. Benchmarking large language models on Bangla dialect translation and dialectal sentiment analysis. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 322–337, Mumbai, India. Association for Computational Linguistics.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2024. BenLLM-eval: A comprehensive evaluation into the potentials and pitfalls of large language models on Bengali NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2238–2252, Torino, Italia. ELRA and ICCL.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4).

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Lawrence I-Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Mehraj Hossain Mahi, Anzir Rahman Khan, and Mayen Uddin Mojumdar. 2025. Bangladial: A merged and imbalanced text dataset for bengali regional dialect analysis. *Data in Brief*, 63:112200.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Jayanta Sadhu, Maneesha Saha, and Rifat Shahriyar. 2024. An empirical study of gendered stereotypes in emotional attributes for Bangla in multilingual large language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 384–398, Bangkok, Thailand. Association for Computational Linguistics.

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2025. Social bias in large language models for Bangla: An empirical study on gender and religious bias. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 204–218, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

K. M. Jubair Sami, Dipto Sumit, Ariyan Hossain, and Farig Sadeque. 2025. A comparative analysis of retrieval-augmented generation techniques for Bengali standard-to-dialect machine translation using LLMs. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 266–279, Mumbai, India. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Azmine Toushik Wasi, Raima Islam, Mst Rafia Islam, Farig Sadeque, Taki Hasan Rafi, and Dong-Kyu Chae. 2025. Dialectal bias in bengali: An evaluation of multilingual large language models across cultural variations. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1380–1384, New York, NY, USA. Association for Computing Machinery.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. 2025. An empirical study of llm-as-a-judge: How design choices impact evaluation reliability. *Preprint*, arXiv:2506.13639.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Translation Fidelity Judge: Full Prompt

The following prompt structure was used for the LLM-as-a-judge translation fidelity evaluation. The judge receives four inputs: the source Bengali sentence, an English gloss, the human reference dialectal translation, and the machine-generated translation. It must complete three structured reasoning steps before returning a JSON response.

Step 1 — Exemptions (No Penalty). The judge is instructed that Bengali dialects lack standardized orthography and that its primary check is *phonetic equivalence*. It must not penalize: (1) phonetic matches: if written forms produce the same or similar dialectal pronunciation (e.g., ধরণ/ধরন, meaning 'type', কালকে/কালকা, meaning 'tomorrow'), they are identical; (2) digit-vs-word number forms (e.g., ৬৪ vs. চয়ষট্টিটা, meaning '64' vs. 'sixty-four'); (3) whitespace and terminal punctuation differences (e.g., ভালা লাগে না vs. ভালালাগেনা, meaning 'does not feel good'); (4) minor dialect-valid morphological suffix variants.

Step 2 — Inaccuracy Count. For differences not exempt under Step 1, the judge counts words falling into two categories: inaccurate_word (wrong dialectal word or incorrect meaning) and meaning_shift (register change such as তুমি vs. আপনি, meaning 'you [informal]' vs. 'you [formal]', or semantic shift such as কিতা vs. কই, meaning 'what' vs. 'where'). A valid dialectal synonym is not counted as an inaccuracy.

Step 3 — Strict Scoring Rubric (0–10).

- **10**: Only exempt differences.
- **9**: Exactly one valid dialectal synonym.
- **8**: One slightly off word; meaning completely preserved.
- **7**: Hard ceiling for exactly one inaccurate word or meaning shift.
- **6**: Exactly two inaccuracies; meaning mostly preserved.

- **5**: Exactly two inaccuracies; meaning noticeably diminished.
- **4**: Three inaccuracies; gist preserved.
- **3**: Three inaccuracies; partially right.
- **1–2**: Four or more inaccuracies, or drastically altered meaning.
- **0**: Complete failure, wrong dialect/language, or hallucination.

JSON Response Format. The judge returns only JSON, executing `chain_of_thought_reasoning` first: (1) read human reference; (2) read machine translation; (3) list exempt phonetic/spacing matches; (4) count remaining inaccurate/shifted words; (5) map to score. The remaining fields are: `exempt_differences_found` (comma-separated list), `inaccurate_words` (comma-separated with brief reason), `meaning_preserved` (`yes`/`partial`/`no`), `score_integer` (integer 0–10), and `score_rationale` (one sentence referencing the rubric and inaccuracy count).

B More Details of the RLAIF Framework

B.1 Confidence Score Guidelines

Judges estimated their probability of correctness on a 1–5 scale based on the following guidelines:

- **Score 5 (Very High / >90% Certainty):** The distinction between responses is obvious; script usage is clear; no cultural nuance ambiguity.
- **Score 4 (High / 75–90% Certainty):** Solid evaluation, but slight nuance might be open to interpretation.
- **Score 3 (Moderate / 50–75% Certainty):** Difficult to interpret dialect (e.g., rare idioms); subjective comparison.
- **Score 2 (Low / 25–50% Certainty):** Significant ambiguity in interpreting Bengali input; lack of specific cultural context.
- **Score 1 (Very Low / <25% Certainty):** Dialect largely unintelligible; responses are gibberish. *Note: If script is indeterminable, Confidence must be 1.*

B.2 Bengali Script Validation

The prompt enforces a critical prerequisite: The response's **primary text** must be written in Bengali script. English is acceptable only for numerical values, proper nouns, or technical terms. If the dialectal response is primarily in Romanized Bengali or another script, all metric scores are automatically set to 0.

B.3 Prompt Structure

The evaluation prompt requires the judge to first generate a `chain_of_thought_reasoning` explicitly comparing the responses before assigning scores, ensuring the quantitative metrics are grounded in qualitative analysis.

C Human Annotators

We recruited 35 native speakers across dialects: Chittagong (8), Sylhet (7), Tangail (5), Rangpur (4), Barishal (1), Noakhali (3), Mymensingh (4), and Kishoreganj (2), plus 1 fallback annotator.

D Evaluated LLMs for Bias Detection

The 19 open-weight LLMs evaluated for dialectal bias detection span the following model families:

- **Gemma**: gemma-3n-e2b, gemma-3n-e4b, gemma-3-12b, gemma-3-27b
- **Llama**: llama-3.1-8b, llama-3.2-3b, llama-3.3-70b
- **Qwen**: qwen3-4b, qwen3-8b, qwen3-32b, qwen-3-235b-a22b-instruct-2507
- **Mistral / Ministral**: mistral-7b, ministral-3-8b, ministral-3-14b
- **DeepSeek**: deepseek-r1-8b, deepseek-r1-32b
- **Phi**: phi4-14b
- **GPT-OSS**: gpt-oss-20b, gpt-oss-120b