Human–LLM Benchmarks for Bangla Dialect Translation: Sylheti & Chittagonian on the BanglaCHQ-Summ Corpus

Nowshin Mahjabin[1,*], Ahmed Shafin Ruhan[1,*], Mehreen Hossain Chowdhury[1,*], Md Fahim[2,3,†], Md. Azam Hossain[1,†]

[1]*Islamic University of Technology*
[2]*Center for Computational & Data Sciences* [3]*Penta Global Limited*
[*]Equal Contribution [†]Project Lead

Correspondence: {nowshin,ahmedshafin,mehreenhossain}@iut-dhaka.edu

Abstract

Millions in Bangladesh speak Sylheti and Chittagonian (Chatgaiyya) dialects, yet most public health guidance exists only in Standard Bangla, which creates barriers and safety risks. Ad-hoc translation further harms comprehension, while challenges such as scarce data, non-standard spelling, medical terms, numerals, and idioms make accurate translation difficult. We present **BanglaCHQ-Prantik**, the first benchmark for this setting, extending BanglaCHQ-Summ with human gold references from 17 native translators. We evaluate Qwen 2.5 3B, Gemma 3 1B, GPT-4o mini, and Gemini 2.5 Flash under zero-shot, one-shot, five-shot, and chain-of-thought prompts, using BLEU, ROUGE-1/2/L, and METEOR. Closed-source models (GPT-4o, Gemini 2.5) lead overall, with Gemini 2.5 Flash being strongest. Few-shot prompting helps especially for Sylheti, though errors persist with terminology, numerals, and idioms. The dataset is designed to support both NLP research and public health communication by enabling reliable translation across regional Bangla dialects. To our knowledge, this is the first medical-domain dataset for Sylheti/Chittagonian.

1 Introduction

Access to clear and reliable health information is critical for safe and effective care. However, in Bangladesh, most official and digital health materials are available only in Standard Bangla (Directorate General of Health Services (DGHS), 2021, 2022; Ministry of Health and Family Welfare (MOHFW), 2024). This poses significant barriers for millions who primarily speak regional dialects such as Sylheti and Chittagonian, spoken by approximately 11 million and 13 million people respectively (syl, 2025; chi, 2025). In the absence of linguistically accessible resources, patients often rely on informal translation or assistance from family members. These practices heighten the risk of miscommunication and medical errors (Al Shamsi et al., 2020). Enabling medical communication in regional dialects is therefore essential to ensure equitable access to health information across the population. Developing accurate dialectal medical translation is very challenging. Parallel resources for Sylheti and Chatgaiyya are scarce, spelling and syntax vary widely, idioms diverge from Standard Bangla and medical queries involve complex terminology, numerals, and dosage expressions. Such factors often lead to translation errors even in strong Machine Translation or Large Language Models (LLMs). While LLMs show promise with few-shot prompting, their ability in low-resource, medical-domain dialectal translation remains underexplored. Previous Bangla NLP studies have focused on related areas such as consumer-health question summarization (BanglaCHQ-Summ) (Khan et al., 2023) and general dialect-to-standard translation (ONUBAD, ChatgaiyyaAlap) (Sultana et al., 2025; Chowdhury et al., 2025) but none address the medical domain or provide benchmarks for dialectal evaluation. We bridge this gap by extending BanglaCHQ-Summ with human-verified Sylheti and Chatgaiyya translations, ensuring semantic fidelity and clinical accuracy. Using this dataset, we benchmark instruction-tuned LLMs Qwen 2.5 3B, Gemma 3 1B, GPT-4o mini, and Gemini 2.5 Flash under zero-shot, one-shot, five-shot, and chain-of-thought prompting, evaluated with BLEU, ROUGE, and METEOR. To clarify the objectives of this work, we position **BanglaCHQ-Prantik** as both a research and practical resource. It is designed to support three primary user communities. First, it provides NLP researchers with a benchmark for studying low-resource dialectal translation and evaluating prompting strategies in Bangla-family languages. Second, it enables machine translation developers to fine-tune and assess large language models or domain-specific translation systems on authentic medical queries. Third, it offers public health com-

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 223–236
December 23, 2025 ©2025 Association for Computational Linguistics

munication practitioners a linguistically grounded tool for producing dialect-sensitive health information and outreach materials. By serving these complementary audiences, **BanglaCHQ-Prantik** bridges the gap between computational research and real-world health communication, reinforcing the dataset's relevance and societal impact.

Our contributions are:

1. We release **BanglaCHQ-Prantik**, a human-validated Sylheti and Chatgaiyya medical translation benchmark with accompanying prompts and scoring scripts.

2. We present the first systematic comparison of open and closed-source LLMs for dialectal medical translation in Bangla, analyzing the effects of prompting strategies.

3. We provide empirical insights into dialectal difficulty, demonstrating that Sylheti is comparatively easier for current large language models (LLMs) than Chittagonian. However, for both dialects, LLM-based translation remains far from accurate or reliable.

2 Dataset Creation

Our aim is to create a new dataset by translating consumer health questions (CHQs) from a Standard Bangla corpus into two major regional dialects: Sylheti and Chittagonian. This process resulted in a parallel corpus where each original data sample is paired with a human-translated version in each dialect. The primary objective is to ensure that the translations are not only linguistically accurate but also preserved all critical clinical information, such as symptoms, durations, and medication details. All human translations have been meticulously reviewed for quality, focusing on accurate terminology and natural phrasing.

2.1 CHQ Dataset Introduction

Our data source is the *BanglaCHQ-Summ* corpus (Khan et al., 2023), which contains original spoken-query passages and their abstractive summaries. This dataset has been chosen because the corpus is situated within the specialized domain of consumer health queries (CHQs). It contains a blend of domain-specific terminology and informal, colloquial expressions, which include conversational elements. By working with this data, we evaluate the models' capacity to not only handle linguistic divergence but also to maintain the fidelity of critical medical information during translation. This presents a more robust and practical challenge than translating standard, well-structured prose.

2.2 Data Filtering and Cleaning

Before starting annotation, we first processed the sourced CHQs to remove irrelevant or inconsistent text while keeping all clinical content intact. **Text Cleaning and Normalization.** We corrected obvious formatting issues that break downstream processing such as: stray characters (i.e. isolated '#' tokens), repeated whitespace, inconsistent punctuation and standardized numerals and measurement units to a consistent textual form (e.g., '5 mg', '5mg' → '5 mg'). We avoided any semantic rewriting. Medical terms, dosages, symptom descriptions, and temporal cues are preserved verbatim except for normalization of punctuation/spacing. **Medical Term Validation.** To validate the preprocessing, two native speakers manually inspected a random sample of 600 entries before and after cleaning. They verified that medical entities, dosages, and question intent were preserved. No loss of clinical information was observed. In these spot checks, we also observed that cleaning consistently fixed tokenization and formatting errors while preserving clinical content and the original question intent/answerability.

2.3 Data Annotation

Nine native speakers produced the reference translations: six for Sylheti and three for Chittagonian. All were native speakers from diverse backgrounds (high school, university graduates, and working adults). Before beginning, annotators received written guidelines with example translations based on four principles: (i) preserve medical meaning, (ii) use idiomatic dialect phrasing, (iii) apply consistent orthography, and (iv) retain terminology, numerals, and units.

The corpus was partitioned by dialect and stratified by length and topic. The 2,350 Sylheti items were split roughly evenly among six annotators, while 500 Chittagonian items were divided among three. Two medical experts independently reviewed the dataset to validate the accuracy of clinical terminology. Translation took 11 days. Annotators gave informed consent, submitted anonymized translations, and were paid Tk. 2 per item. Annotators self-reviewed their work and the study team ran random spot checks, returning flagged items for

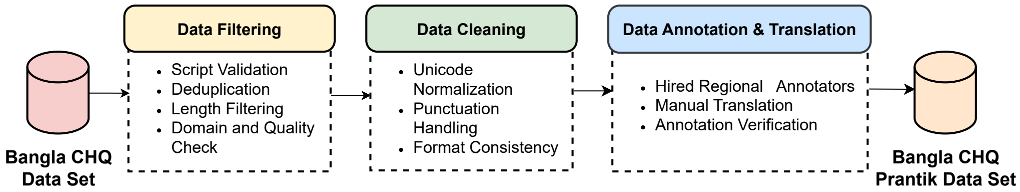


Figure 1: Data processing pipeline for constructing the Bangla CHQ Prantik dataset. The raw Bangla CHQ data undergoes data filtering (script validation, deduplication, length filtering, domain checks), followed by data cleaning (Unicode normalization, punctuation handling, format consistency), and finally data annotation and translation by regional annotators before producing the finalized dataset.

minor corrections.

2.4 Data Validation

We employed eleven independent validators to ensure the quality and accuracy of our translations. Given that our parent dataset BangCHQ already contained validated medical terminology, our emphasis was on recruiting validators with diverse linguistic backgrounds who could assess the naturalness and appropriateness of regional language use, rather than primarily focusing on medical expertise.

Sylheti. The 2,350 items were divided into two equal halves. Validators A and B jointly reviewed the first 1,175 items, while Validators C and D reviewed the second half. A fifth validator (E) served as adjudicator for any disagreements. Inter-validator agreement was assessed using BLEU, BERTScore, METEOR, and ROUGE-L (F1) metrics. For Validators A & B, the scores were 73.28%, 95.24%, 82.77%, and 86.43%, respectively; for C & D, the scores were 71.12%, 93.49%, 80.26%, and 82.57%.

Chittagonian. Validators E and F independently reviewed all 500 items. A third validator adjudicated disagreements, with the majority vote determining the final version. Agreement was again measured using BLEU, BERTScore, METEOR, and ROUGE-L (F1), yielding scores of 70.97%, 94.30%, 79.94%, and 81.18%, respectively.

Expert Medical Validation. To ensure clinical accuracy, we conducted an additional expert validation phase with two licensed physicians who independently reviewed a random sample of 75 items from the combined dataset. The medical experts assessed the preservation of clinical intent and the appropriateness of translated medical terminology in regional contexts. Inter-annotator agreement between the two physicians, measured using BLEU, BERTScore, METEOR, and ROUGE-L (F1), yielded scores of 89.45%, 97.12%, 91.33%, and 93.76%, respectively, indicating very high consistency in their clinical assessments.

Validators reviewed spelling, numerals, clinical terminology, and fidelity to the original question's intent. While automatic similarity metrics offered rough signals of agreement, all final decisions were human-led. The full validation process took seven days. Appendix tables include flagged items, representative disagreements, and their resolutions.

3 Dataset Statistics

Statistics	Bangla CHQ	Ours	
		Sylheti	Chittagonian
Mean Char. Length	325.71	323.24	310.84
Max Char. Length	868	881	531
Min Char. Length	225	189	181
Mean Word Count	67.50	60.73	57.84
Max Word Count	169	166	93
Min Word Count	39	31	36
#Unique Words	12447	18399	2628
#Unique Sentence	12136	14145	796

Table 1: Dataset statistics of the Bangla CHQ and Sylheti CHQ columns.

The dataset statistics for the proposed **BANGLA CHQ PRANTIK** resource are summarized in Table1. It includes the full set of 2,350 CHQ items translated into Sylheti and a 500-item subset translated into Chittagonian.

The Chittagonian portion is substantially smaller due to practical and linguistic challenges we encountered during data collection. The primary bottleneck was the scarcity of annotators proficient in reading and writing Chittagonian. Unlike Sylheti, which has a more established written tradition and a larger pool of literate speakers, Chittagonian lacks a widely accepted standard orthography and has limited written materials. This has resulted in fewer speakers with the literacy skills needed for specialized translation work, particularly in healthcare. Additionally, the dialect varies considerably across the Chittagong division, requiring annotators to make careful decisions about which forms would be most broadly understood. These factors added substantial time to each translation, as spelling, terminology, and phrasing choices required more deliberation and validation than was typical for Sylheti.

Given these challenges, we prioritized translation accuracy and dialectal authenticity over quantity, recognizing that a smaller, rigorously validated dataset would be more valuable than a larger corpus of questionable quality. Despite the smaller sample size, the Chittagonian subset still manages to capture essential dialectal variation and serves as an important step toward broader linguistic inclusivity in Bangla health resources. Future work could expand the Chittagonian dataset through several practical steps. Developing clearer orthographic guidelines in collaboration with linguistic experts would give annotators a more consistent framework to work from. A phased expansion approach could also prove effective, where new translations go through community review to validate quality while simultaneously training additional annotators. Machine translation may eventually help with initial drafts, though given the dialect's complexity, human oversight will remain essential for the foreseeable future.

4 Experiment Setup

Zero-Shot Prompting. To evaluate the capabilities of LLMs, we first employ the *zero-shot prompting* approach. Each model is provided with a system prompt P and a dialect text T_{Dialect}, and is tasked with generating the corresponding formal text T_{Formal} without access to any example pairs (Brown et al., 2020).

Few-Shot Prompting. In the *few-shot prompting* setting, in addition to the system prompt P and input dialect text T_{Dialect}, the model is also given a set of k labeled example pairs $\mathbf{E} = (T_D^1, T_F^1), \dots, (T_D^k, T_F^k)$, where each (T_D^i, T_F^i) pair represents a dialect sentence and its corresponding formal version (Brown et al., 2020).

Chain-of-Thought (CoT) Prompting. For *Chain-of-Thought (CoT)* prompting, we guide the model to engage in intermediate reasoning before generating a response (Wei et al., 2022). This is achieved by appending the phrase *"Let's think step by step"* to the system prompt P, encouraging the model to produce multi-step inferences leading up to the final output (Kojima et al., 2022). In our translation setting, we further instructed the model to analyze the text step by step before translating, prompting it to reason about meaning, structure, and terminology prior to producing the final translation. Details of the prompt configurations are provided in the Appendix D.

5 Result and Analysis

Model Performance. Closed-source systems (GPT-4o, Gemini 2.5) outperform open-source baselines (Qwen 2.5 3B, Gemma 3 1B), reflecting advantages in scale, alignment, and multilingual corpora. Gemini 2.5 is strongest especially on Sylheti with 5-shot prompting (BLEU 23.67, ROUGE-1 49.02, METEOR 43.82) likely due to broader linguistic coverage and more effective Bangla tokenization. GPT-4o is competitive but slightly lower; Gemma 3 1B surpasses Qwen 2.5 3B, while both open models show limited exposure to dialectal Bangla.

Prompting Impact. Few-shot prompting yields the most reliable gains, notably for Gemini 2.5, by supplying lexical/morphological anchors for medical phrasing. Chain-of-thought (CoT) shows smaller, less stable gains, sometimes falling below both 1-shot and 5-shot, likely because explicit reasoning overgeneralizes Sylheti and Chittagonian's repetitive syllables and reduces lexical fidelity.

Dialect-wise Analysis. Sylheti consistently outperforms Chittagonian across models. Even with CoT, Chittagonian lags (BLEU 21.53 *vs.* Sylheti ROUGE-1 33.37), showing greater differences in pronunciation and structure compared to Standard Bangla, sparser pretraining exposure, and less standardized orthography. Sylheti benefits from wider representation in online and digital resources, contributing to more stable spelling conventions.

Error Analysis. The main errors are (i) terminology mismatches or omissions, (ii) numeral/unit mistakes, (iii) literalized idioms, and (iv) orthographic drift. Closed-source models better preserve medical terms, whereas open models tend to translate too literally, often missing contextual nuances and producing less adequate outputs. Few-shot examples reduce numeric and lexical issues but leave

Model Name	Sylheti					Chittagonian				
	B	R1	R2	R_L	Met	B	R1	R2	R_L	Met
					Zero Shot Prompt					
Open Source										
Qwen 2.5 3B	9.4	32.97	14.76	31.18	22.47	3.35	16.83	4.2	15.71	10.15
Gemma 3 1B	12.64	33.84	12.41	32.13	25.84	6.93	24.04	6.99	23.3	18.44
Closed Source										
GPT-4o	22.9	46.36	22.98	45.61	41.55	9.08	27.74	8.76	27.46	22.86
Gemini 2.5	20.97	46.33	22.56	45.4	41.09	14.66	37.12	15.61	37.06	32.86
					1 Shot Prompt					
Open Source										
Qwen 2.5 3B	8.68	29.10	13.01	27.50	19.99	3.12	15.43	3.75	14.53	9.27
Gemma 3 1B	13.52	38.17	15.87	35.18	33.29	5.05	19.89	5.67	18.32	16.43
Closed Source										
GPT-4o	22.23	45.95	22.32	45.19	40.92	9.02	28.04	8.94	27.81	23.02
Gemini 2.5	22.19	47.49	23.82	46.64	42.34	11.56	33.61	12.85	33.41	28.99
					5 Shot Prompt					
Open Source										
Qwen 2.5 3B	8.57	31.38	13.61	29.28	21.11	3.36	16.35	4.06	15.21	9.91
Gemma 3 1B	17.24	40.29	17.48	38.78	33.44	7.22	23.97	7.21	23.51	19.1
Closed Source										
GPT-4o	22.34	46.28	22.42	45.38	40.78	8.99	29.07	9.11	28.67	23.44
Gemini 2.5	23.67	49.02	25.2	49.97	43.82	14.78	36.76	14.95	38.85	34.96
					CoT Prompt					
Open Source										
Qwen 2.5 3B	7.16	29.41	11.66	26.99	18.93	3.36	16.3	4.06	15.21	9.91
Gemma 3 1B	23.57	47.24	24.88	46.53	43.6	7.34	23.8	6.28	22.74	18.9
Closed Source										
GPT-4o	21.8	45.81	21.6	44.72	39.71	24.67	51.74	26.84	50.57	47.3
Gemini 2.5	21.53	46.84	22.86	46.02	41.54	9.29	33.37	12.1	32.3	28.07

Table 2: Model benchmarking results on the test split of the BANGLA CHQ PRANTIK dataset across four prompting strategies. B, R1, R2, R_L, and Met represent BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR scores, respectively. Blue text indicates the highest-performing model for each metric within each prompting strategy configuration (Zero Shot, 1 Shot, 5 Shot, and CoT).

idiomatic and orthographic inconsistencies largely unresolved (see Appendix C).

6 Conclusion

We present **BANGLA CHQ PRANTIK** , extending BanglaCHQ-Summ with human-validated translations for two major Bangla dialects—*Sylheti* and *Chittagonian*. Across zero-shot, few-shot, and CoT settings, five LLMs were evaluated with BLEU, ROUGE, and METEOR; closed-source models (notably Gemini 2.5) consistently outperformed open-source baselines, and Sylheti proved easier than Chittagonian. Persistent errors involved medical terminology, numerals/units, and idioms, underscoring the need for richer dialectal resources. Future work will extend BANGLA CHQ PRANTIK beyond text to *synthetic audio* and *speech-to-text*: developing dialectal TTS for ASR, investigating code-switching and mixed-script cases (Bangla–English, Romanized Bangla), constructing multi-reference test sets, and scaling human evaluation of adequacy, fluency, and dialect authenticity (Khan et al., 2023). For healthcare applications, priorities include robustness to spoken input and user-centered evaluation to ensure accuracy, safety, and usability.

Limitations

Our work has several limitations. First, evaluation relies on single-reference translations, which penalize legitimate dialectal variation in spelling or phrasing. Second, closed-source systems are black-boxes, limiting insight into their training data or dialectal exposure. Third, automatic metrics (BLEU, ROUGE, METEOR) cannot fully capture the clinical adequacy of translations; a human adequacy/fluency study would strengthen conclusions. Finally, we focus exclusively on text translation; extending to spoken dialects and multimodal contexts is an important future step.

Ethics Statement

BanglaCHQ-Prantik is developed entirely from anonymized consumer health queries sourced from the publicly available BanglaCHQ-Summ corpus. The dataset contains no personally identifiable information, and no such information was collected, stored, or distributed at any stage of the project. All dialectal translations were produced by native speakers who provided informed consent and received fair compensation for their work; no sensitive personal or demographic details about annotators are included. The dataset focuses solely on linguistic variation within medical-domain text and does not contain material intended to harm, stigmatize, or misrepresent any individual or community. Although the corpus involves health-related content, it is not designed or intended for clinical decision-making or the provision of medical advice. Based on these considerations, we do not anticipate any ethical concerns associated with the release or use of BanglaCHQ-Prantik.

Acknowledgements

Our work is supported by the Islamic University of Technology Research Seed Grants (IUT RSG) (Ref: REASP/IUT-RSG/2022/OL/07/012).

References

2025. Chittagonian (chit1275) — glottolog 5.2. https://glottolog.org/resource/languoid/id/chit1275. Accessed: 2025-10-04.

2025. Sylheti (sylh1242) — glottolog 5.2. https://glottolog.org/resource/languoid/id/sylh1242. Accessed: 2025-10-04.

Hilal Al Shamsi, Abdullah G. Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for healthcare: A systematic review. *Oman Medical Journal*, 35(2):e122.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (MTEval 2005)*, pages 65–72.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sinthia Chowdhury, Deawan Rakin Ahamed Remal, Sheak Rashed Haider Noori, Syed Tangim Pasha, and Ashraful Islam. 2025. Chatgaiyyaalap: A dataset for conversion from chittagonian dialect to standard bangla. *Data in Brief*, 59:111413. Dataset version: Mendeley Data V1, DOI:10.17632/wtms9xbkkw.1.

Directorate General of Health Services (DGHS). 2021. Covid-19 (bangladesh covid-19 management guideline). https://dghs.gov.bd/site/page/86973b0c-62dd-4a43-92f2-b305e3264c88/. Official public health guideline published in Standard Bangla by the Directorate General of Health Services, Ministry of Health and Family Welfare, Bangladesh.

Directorate General of Health Services (DGHS). 2022. Mental health and covid-19 guideline. https://dghs.gov.bd/site/page/86973b0c-62dd-4a43-92f2-b305e3264c88/. Standard Bangla document under the National Health Communication Program, Bangladesh.

Alvi Khan, Fida Kamal, Mohammad Abrar Chowdhury, Tasnim Ahmed, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. BanglaCHQ-summ: An abstractive summarization dataset for medical queries in Bangla conversational speech. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 85–93, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Ministry of Health and Family Welfare (MOHFW). 2024. Ministry of health and family welfare, government of bangladesh: Official portal (in standard bangla). https://mohfw.gov.bd/. Official Bangla-language public health resources, circulars, and program descriptions.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT): Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. Introduces sacreBLEU for standardized BLEU reporting.

Nusrat Sultana, Rumana Yasmin, Bijon Mallik, and Mohammad Shorif Uddin. 2025. Onubad: A comprehensive dataset for automated conversion of bangla regional dialects into standard bengali dialect. *Data in Brief*, 58:111276. Dataset article; freely accessible via Mendeley at https://data.mendeley.com/datasets/6ft99kf89b/2.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

A Dataset Comparison

We showed a comparison between different dialects of a single text in our Bangla CHQ Prantik dataset. The comparison is shown in the Figure 2, with English translation provided in the diagram to aid comprehension. This figure illustrates dialectal variation in medical-domain text, showing how a single health-related message in English is translated into Standard Bangla, Chittagonian, and Sylheti. While the core meaning is preserved across versions, there are notable differences in vocabulary, morphology, and phrasing. For instance, the word for "ear" appears in the three variants, and verbs such as "hurt" and "came out" are expressed using distinct regional forms.

These differences highlight the linguistic distance between Standard Bangla and regional dialects, which often lack standardized orthographies and are not mutually intelligible. Such variation poses challenges for both human understanding and machine translation in healthcare settings. This example underscores the importance of dialect-specific resources—such as BANGLA CHQ PRANTIK ensure equitable access to medical information for all language communities in Bangladesh.

B Experiment Details

Gemma Settings: Experiments on Gemma were conducted on Kaggle Notebooks with Python 3.11.13. Hardware resources consisted of two NVIDIA Tesla T4 GPUs (16 GB VRAM each), 4 vCPUs, and 31 GB system RAM. The model was executed with `float32` precision and GPU memory utilization set to 0.7, using tensor parallelism across two GPUs. The maximum model length was configured to 2048 tokens, with up to 24 parallel sequences.

Qwen Settings: Experiments on Qwen2.5 3B were conducted on Kaggle Notebooks with Python 3.11.13. Hardware resources included an NVIDIA Tesla T4 GPU (16 GB VRAM) alongside 4 vCPUs and 31 GB of system RAM. The model was run with `float16` precision and GPU memory utilization set to 0.65. The maximum model length was configured to 2048 tokens, with up to 24 parallel sequences.

Gemini: For Gemini, the gemini-2.5-flash model was accessed through the official API. All inference requests were handled remotely, following the platform's default settings.

GPT: Experiments with GPT used the GPT-4o-mini model via the official OpenAI API. Inference was performed with default API settings.

B.1 Evaluation Metrics

We evaluated system outputs against human references using three standard MT families: **BLEU**, **ROUGE** (1/2/L), and **METEOR**. For each translation pair we compute the metrics below, then report corpus-level statistics (mean, median, min/max, std) for the main dataset. Results are shown separately for Sylheti and Chittagonian under 0-shot, 5-shot, and CoT settings.

BLEU (higher is better): We report corpus-level BLEU using **sacreBLEU** (Post, 2018), which provides a standardized evaluation pipeline and generates a unique signature to ensure exact reproducibility. For Bengali-script text, we use sacreBLEU's built-in tokenization (with no custom preprocessing) to ensure consistency. Final scores are averaged at the corpus level and presented as percentages in column **B**.

ROUGE-1/2/L (higher is better): We compute ROUGE-1 and ROUGE-2 (unigram/bigram overlap) and ROUGE-L (longest common subsequence) using the official `rouge_score` implementation (Lin, 2004). We report the F_1 variant without stemming, again as percentages (columns **R1**, **R2**, **R_L**).

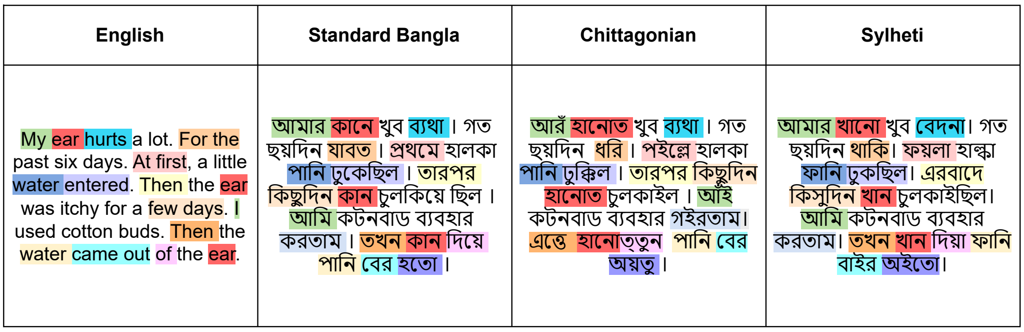
METEOR (higher is better): METEOR (Banerjee and Lavie, 2005) is computed with NLTK's implementation over whitespace tokens (no stemming or paraphrase tables), yielding a precision/recall–balanced score. We report percentages in the **Met** column. METEOR rewards partial matches and recall, making it more sensitive to meaning preservation and synonymy—particularly valuable for dialectal Bangla, where lexical variation and spelling differences are common.

Interpretation: BLEU emphasizes n-gram precision with a brevity penalty; ROUGE captures lexical overlap and sequence alignment; METEOR balances precision and recall with alignment heuristics. These metrics complement each other by reflecting both word-level accuracy and sequence-level similarity in dialectal translations.

C Extended Error Analysis and Findings

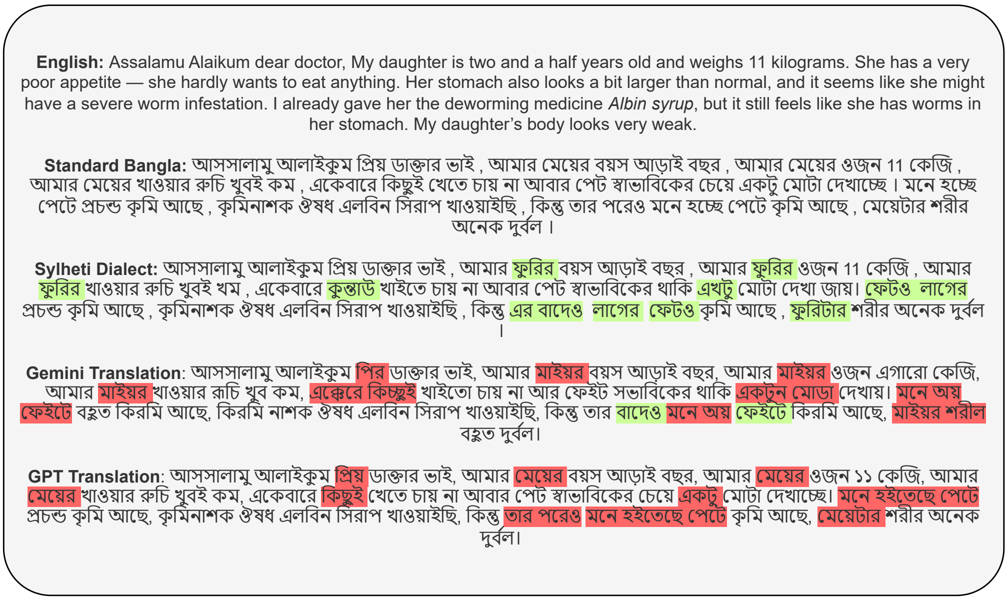
C.1 Model Performance

Closed-source models (GPT-4o, Gemini 2.5) consistently outperform open-source ones across all settings. Gemini 2.5 shows particular robustness in Sylheti, leveraging in-context learning to achieve state-of-the-art results. GPT-4o remains strong across both dialects, though generally a step behind. Among open-source models, Gemma 3 1B consistently outperforms Qwen 2.5 3B, reflecting stronger instruction tuning. However, both open



English	Standard Bangla	Chittagonian	Sylheti
My ear hurts a lot. For the past six days. At first, a little water entered. Then the ear was itchy for a few days. I used cotton buds. Then the water came out of the ear.	আমার কানে খুব ব্যথা । গত ছয়দিন যাবত । প্রথমে হালকা পানি ঢুকেছিল । তারপর কিছুদিন কান চুলকিয়ে ছিল । আমি কটনবাড ব্যবহার করতাম । তখন কান দিয়ে পানি বের হতো ।	আরঁ হানোত খুব ব্যথা । গত ছয়দিন ধরি । পইল্লে হালকা পানি ঢুক্কিল । তারপর কিছুদিন হানোত চুলকাইল । আঁই কটনবাড ব্যবহার গইরতাম। এত্তে হানোত্তুন পানি বের অয়তু ।	আমার খানো খুব বেদনা। গত ছয়দিন থাকি। ফয়লা হাল্কা ফানি ঢুকছিল। এরবাদে কিসুদিন খান চুলকাইছিল। আমি কটনবাড ব্যবহার করতাম। তখন খান দিয়া ফানি বাইর অইতো।

Figure 2: Demonstration of dialectal variation in medical-domain text. A Standard Bangla sentence (left) is shown alongside its Chittagonian (middle) and Sylheti (right) translations. Highlighted segments illustrate differences in word choice, morphology, and phrasing across dialects.



English: Assalamu Alaikum dear doctor, My daughter is two and a half years old and weighs 11 kilograms. She has a very poor appetite — she hardly wants to eat anything. Her stomach also looks a bit larger than normal, and it seems like she might have a severe worm infestation. I already gave her the deworming medicine *Albin syrup*, but it still feels like she has worms in her stomach. My daughter's body looks very weak.

Standard Bangla: আসসালামু আলাইকুম প্রিয় ডাক্তার ভাই , আমার মেয়ের বয়স আড়াই বছর , আমার মেয়ের ওজন 11 কেজি , আমার মেয়ের খাওয়ার রুচি খুবই কম , একেবারে কিছুই খেতে চায় না আবার পেট স্বাভাবিকের চেয়ে একটু মোটা দেখাচ্ছে । মনে হচ্ছে পেটে প্রচন্ড কৃমি আছে , কৃমিনাশক ঔষধ এলবিন সিরাপ খাওয়াইছি , কিন্তু তার পরেও মনে হচ্ছে পেটে কৃমি আছে , মেয়েটার শরীর অনেক দুর্বল ।

Sylheti Dialect: আসসালামু আলাইকুম প্রিয় ডাক্তার ভাই , আমার ফুরির বয়স আড়াই বছর , আমার ফুরির ওজন 11 কেজি , আমার ফুরির খাওয়ার রুচি খুবই খম , একেবারে কুন্তাউ খাইতে চায় না আবার পেট স্বাভাবিকের থাকি এখটু মোটা দেখা জায়। ফেটও লাগের প্রচন্ড কৃমি আছে , কৃমিনাশক ঔষধ এলবিন সিরাপ খাওয়াইছি , কিন্তু এর বাদেও লাগের ফেটও কৃমি আছে , ফুরিটার শরীর অনেক দুর্বল ।

Gemini Translation: আসসালামু আলাইকুম পির ডাক্তার ভাই, আমার মাইয়র বয়স আড়াই বছর, আমার মাইয়র ওজন এগারো কেজি, আমার মাইয়র খাওয়ার রূচি খুব কম, এক্কেরে কিচ্ছুই খাইতো চায় না আর ফেইট সভাবিকের থাকি একটুন মোডা দেখায়। মনে অয় ফেইটে বহুত কিরমি আছে, কিরমি নাশক ঔষধ এলবিন সিরাপ খাওয়াইছি, কিন্তু তার বাদেও মনে অয় ফেইটে কিরমি আছে, মাইয়র শরীল বহুত দুর্বল।

GPT Translation: আসসালামু আলাইকুম প্রিয় ডাক্তার ভাই, আমার মেয়ের বয়স আড়াই বছর, আমার মেয়ের ওজন ১১ কেজি, আমার মেয়ের খাওয়ার রুচি খুবই কম, একেবারে কিছুই খেতে চায় না আবার পেট স্বাভাবিকের চেয়ে একটু মোটা দেখাচ্ছে। মনে হইতেছে পেটে প্রচন্ড কৃমি আছে, কৃমিনাশক ঔষধ এলবিন সিরাপ খাওয়াইছি, কিন্তু তার পরেও মনে হইতেছে পেটে কৃমি আছে, মেয়েটার শরীর অনেক দুর্বল।

Figure 3: Error analysis across dialectal outputs produced by different models.

models struggle with dialectal adaptation, producing literal or fragmented outputs.

C.2 Prompting Impact

Few-shot prompting provides the largest improvements, particularly for Gemini 2.5, which benefits from contextual anchors that guide morphological and lexical choices. CoT prompting, while theoretically promising, offers only modest improvements and sometimes underperforms compared to 5-shot. This suggests that reasoning-based approaches are less effective for low-resource dialectal translation than concrete in-context examples.

C.3 Dialect-wise Analysis

Sylheti translations consistently achieve higher scores than Chittagonian. For example, even under CoT prompting, Chittagonian lags (BLEU 21.53 vs. Sylheti ROUGE-1 33.37). This disparity reflects two main factors: (i) dialectal distance, as Chittagonian diverges more from Standard Bangla in phonology and morphosyntax; and (ii) data scarcity, since Chittagonian is less likely to appear in large-scale pretraining corpora. Orthographic variability also increases inconsistency, especially in open-source outputs.

C.4 Case Study: Medical Domain Translation Challenges

To illustrate the error patterns observed across models, we present a representative example from the medical domain. The source sentence describes a pediatric case involving deworming medication ("It seems there are worms in the stomach, I gave

Alben syrup deworming medicine, but still it seems there are worms in the stomach, the girl's body is very weak").

The reference Sylheti translation uses distinct dialectal features: *feto* (stomach) instead of Standard Bangla *pete*, *lager* (seems) for *mone hochhe*, and *furita* (girl) for *meyeta*. Both Gemini and GPT translations exhibit characteristic errors. The Gemini output introduces phonological overcorrections (*feite* instead of *feto*, *kirimi* for *krimi*), demonstrating inconsistent application of dialectal phonology rules. The GPT translation maintains closer orthographic fidelity to Standard Bangla (*pete*, *meyetar*), failing to properly dialectalize key lexical items. Notably, both models preserve the medical term "Alben syrup" correctly, supporting our finding that closed-source models handle domain-specific terminology more reliably. However, dialectal function words show variation: the reference uses *er badeo* (after this), while GPT retains the Standard form *tar poreo*, and Gemini substitutes *tar badeo*—a hybrid form. This example encapsulates the tension between lexical preservation and dialectal authenticity that characterizes medical translation in low-resource settings.

C.5 Error Analysis

We identify four dominant error categories:

1. **Terminology mismatches**: Medical terms (e.g., "antibiotic", "hypertension") are often replaced with generic synonyms or omitted altogether. Closed-source models handle these terms more reliably.

2. **Numerals and units**: Dosages, dates, and measurements are inconsistently translated. Example: "500mg" incorrectly rendered as "5 gram" in Qwen outputs.

3. **Idiomatic expressions**: Dialect-specific idioms are frequently literalized. For instance, Sylheti expressions for pain severity were mapped word-for-word rather than into natural phrasing.

4. **Orthographic drift**: Particularly in Chittagonian, spelling variation leads to inconsistent forms across outputs. This problem is amplified in open-source models.

Few-shot prompting reduces terminology and numeral errors by providing concrete examples. CoT offers some benefit for coverage but does not resolve idiomatic or orthographic inconsistencies.

D Used Prompts in the Paper

Prompt for Bangla to Sylheti Translation (ZeroShot)

ZeroShot Prompt for Bangla to Sylheti Translation

You are a precise translation tool. Your only task is to translate the given Bangla text to Sylheti dialect using Bengali script.

INSTRUCTIONS:

- Translate the Bangla text below to Sylheti dialect
- Use only Bengali script (not Latin script or IPA)
- Return ONLY the translated text with no additional commentary, explanations, or notes
- Do not include phrases like "Here is the translation:" or "The Sylheti translation is:"
- Do not add any metadata, formatting, or extra information
- If you cannot translate a specific word, keep it as is in the original form

Bangla text to translate: `{bangla_text}`

Sylheti translation:

Prompt for Bangla to Sylheti Translation (FewShot)

FewShot Prompt for Bangla to Sylheti Translation

You are a precise translation tool. Your only task is to translate the given Bangla text to Sylheti dialect using Bengali script.

INSTRUCTIONS:

- Translate the Bangla text below to Sylheti dialect
- Use only Bengali script (not Latin script or IPA)
- Return ONLY the translated text with no additional commentary, explanations, or notes
- Do not include phrases like "Here is the translation:" or "The Sylheti translation is:"
- Do not add any metadata, formatting, or extra information
- If you cannot translate a specific word, keep it as is in the original form

Here are some examples of Bangla to Sylheti translations:

Bangla: "Do you meditate regularly?" | Sylheti: [Sylheti translation]
Bangla: "Where do you do coaching?" | Sylheti: [Sylheti translation]
Bangla: "Has the lentil been cooked?" | Sylheti: [Sylheti translation]
Bangla: "Hey there, what are you taking?" | Sylheti: [Sylheti translation]
Bangla: "Will go after prayer" | Sylheti: [Sylheti translation]

Bangla text to translate: `{bangla_text}`

Sylheti translation:

Prompt for Bangla to Sylheti Translation (Chain-of-Thought)

Chain-of-Thought Prompt for Bangla to Sylheti Translation

You are a precise translation tool. Your task is to translate the given Bangla text to Sylheti dialect using Bengali script.

INSTRUCTIONS:

- Analyze the text step-by-step before translating
- Use only Bengali script in your final translation
- After your reasoning, provide ONLY the final translation with no additional commentary

Here are examples showing the translation process:

Ex 1: Bangla: “Do you meditate regularly?” | Reasoning: Phonetic shifts, verb changes | Sylheti: [Translation]

Ex 2: Bangla: “Where do you do coaching?” | Reasoning: Locative form, verb change | Sylheti: [Translation]

Ex 3: Bangla: “Has the lentil been cooked?” | Reasoning: Vowel shift, verb transformation | Sylheti: [Translation]

Ex 4: Bangla: “Hey there, what are you taking?” | Reasoning: Colloquial address, particle transformation | Sylheti: [Translation]

Ex 5: Bangla: “Will go after prayer” | Reasoning: Arabic loanword, future verb ending | Sylheti: [Translation]

Now translate: `{bangla_text}` | **Lets think step by step:** Identify transformations, grammatical changes, phonetic patterns

Provide your final translation below: Sylheti:

Prompt for Bangla to Chittagonian Translation (ZeroShot)

ZeroShot Prompt for Bangla to Chittagonian Translation

You are a precise translation tool. Your only task is to translate the given Bangla text to Chittagonian (Chatgaiyan) dialect using Bengali script.

INSTRUCTIONS:

- Translate the Bangla text below to Chittagonian (Chatgaiyan) dialect
- Use only Bengali script (not Latin script or IPA)
- Return ONLY the translated text with no additional commentary, explanations, or notes
- Do not include phrases like “Here is the translation:” or “The Chittagonian translation is:”
- Do not add any metadata, formatting, or extra information
- If you cannot translate a specific word, keep it as is in the original form

Bangla text to translate: `{bangla_text}`

Chittagonian translation:

Prompt for Bangla to Chittagonian Translation (FewShot)

FewShot Prompt for Bangla to Chittagonian Translation

You are a precise translation tool. Your only task is to translate the given Bangla text to Chittagonian (Chatgaiyan) dialect using Bengali script.

INSTRUCTIONS:

- Translate the Bangla text below to Chittagonian (Chatgaiyan) dialect
- Use only Bengali script (not Latin script or IPA)
- Return ONLY the translated text with no additional commentary, explanations, or notes
- Do not include phrases like “Here is the translation:” or “The Chittagonian translation is:”
- Do not add any metadata, formatting, or extra information
- If you cannot translate a specific word, keep it as is in the original form

Here are some examples of Bangla to Chittagonian translations:

Bangla: “Uncle, will you go to the village house?” | Chittagonian: [Translation]
Bangla: “Do you regularly eat vegetables?” | Chittagonian: [Translation]
Bangla: “Will go after noon prayer” | Chittagonian: [Translation]
Bangla: “Really like spinach” | Chittagonian: [Translation]
Bangla: “What do the maternal grandparents of Rangani do?” | Chittagonian: [Translation]

Bangla text to translate: `{bangla_text}`

Chittagonian translation:

Prompt for Bangla to Chittagonian Translation (Chain-of-Thought)

Chain-of-Thought Prompt for Bangla to Chittagonian Translation

You are a precise translation tool. Your task is to translate the given Bangla text to Chittagonian (Chatgaiyan) dialect using Bengali script.

INSTRUCTIONS:

- Analyze the text step-by-step before translating
- Use only Bengali script in your final translation
- After your reasoning, provide ONLY the final translation with no additional commentary

Here are examples showing the translation process:

Ex 1: Bangla: “Uncle, will you go to the village?” | Reasoning: Vocative particle, possessive/locative | Chittagonian: [Translation]

Ex 2: Bangla: “Do you regularly eat vegetables?” | Reasoning: Pronoun shift, habitual form | Chittagonian: [Translation]

Ex 3: Bangla: “Will go after noon prayer” | Reasoning: Arabic loanword, future verb | Chittagonian: [Translation]

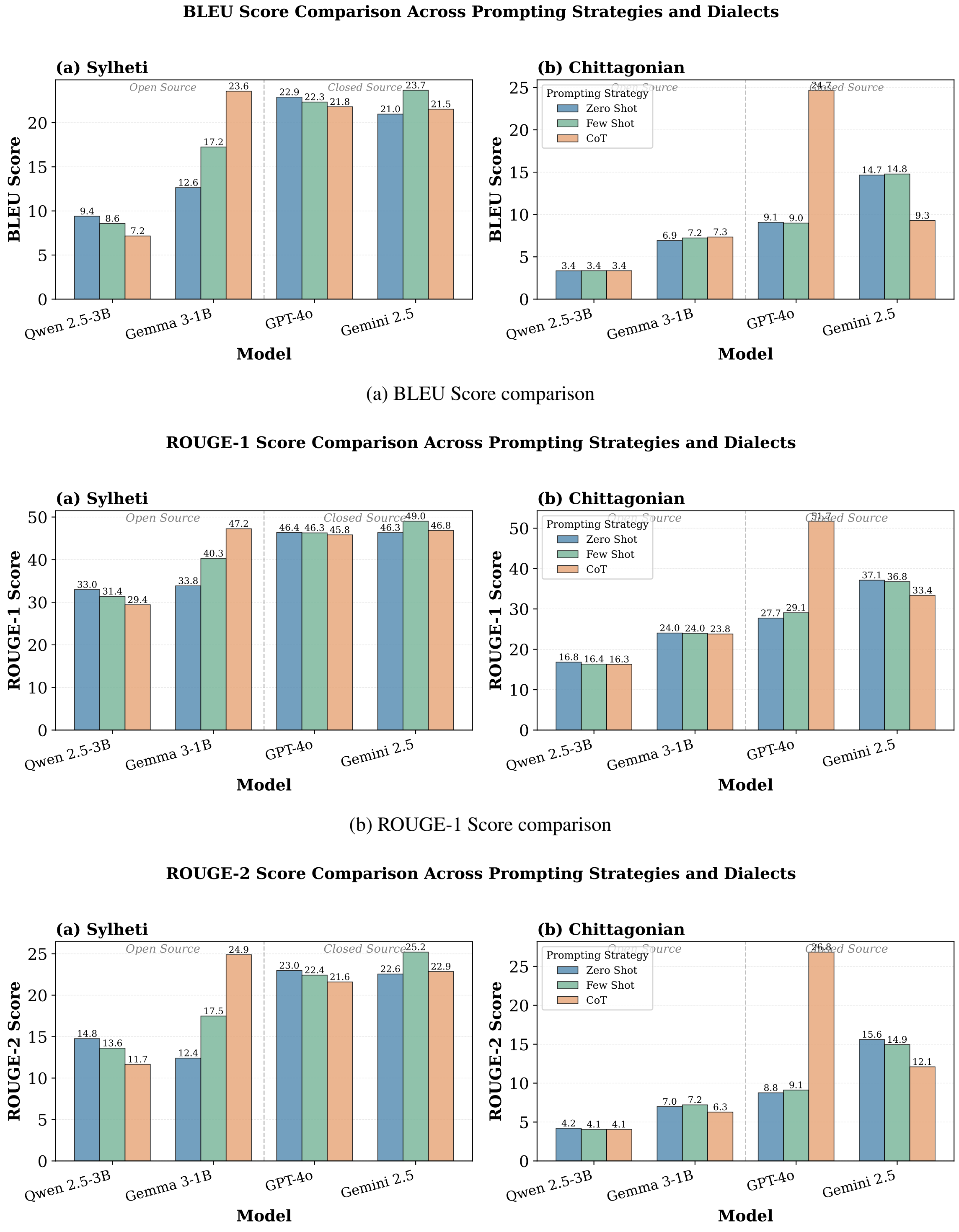
Ex 4: Bangla: “Really like spinach” | Reasoning: Compound unchanged, preference transformation | Chittagonian: [Translation]

Ex 5: Bangla: “What do maternal grandparents do?” | Reasoning: Phonetic compression, plural transformation | Chittagonian: [Translation]

Now translate: `{bangla_text}` | **Let’s think step by step:** Identify transformations, grammatical changes, phonetic patterns

Provide your final translation below: Chittagonian:

E Performance Comparison across Dialects and Prompting Strategies in Different Models

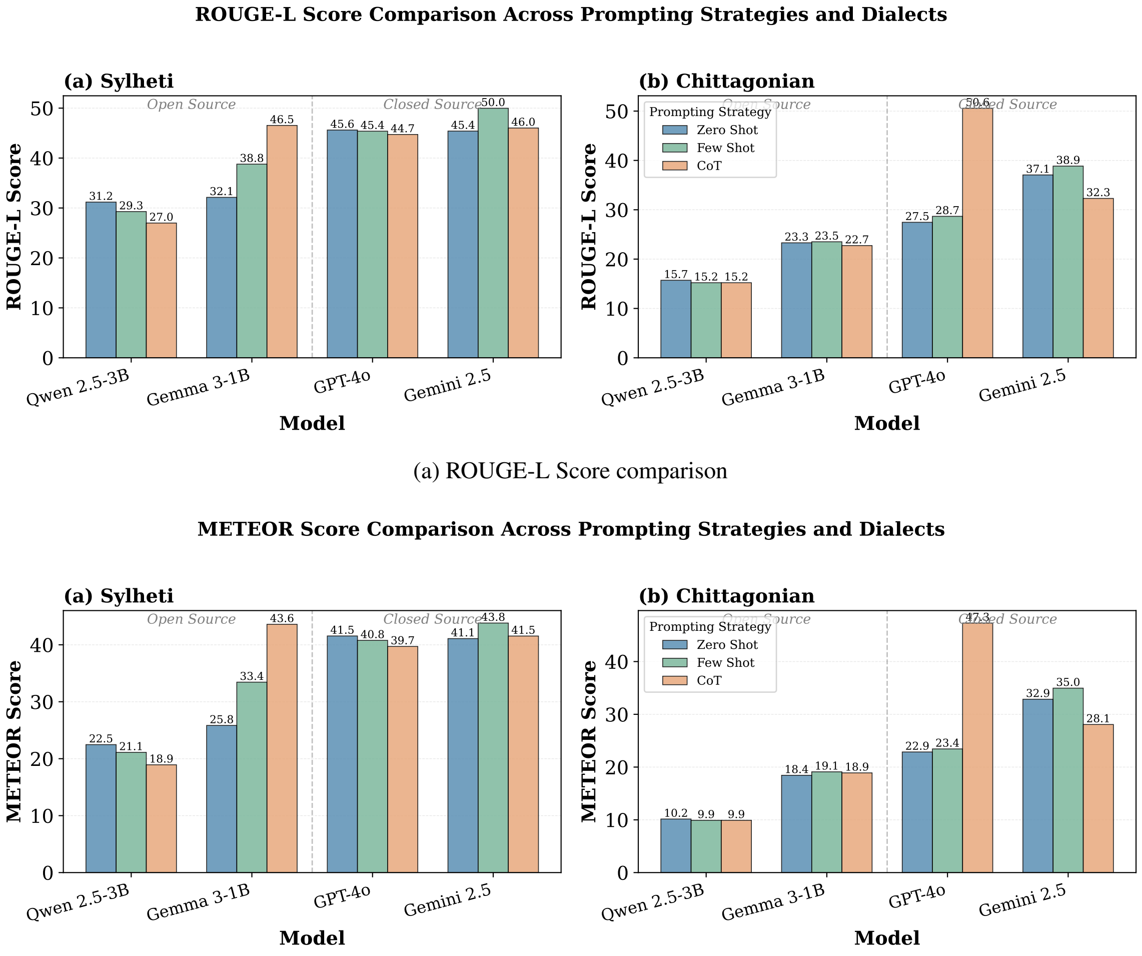


(a) BLEU Score comparison

(b) ROUGE-1 Score comparison

(c) ROUGE-2 Score comparison

Figure 4: Performance comparison across different prompting strategies (Part 1). Comparison of BLEU, ROUGE-1, and ROUGE-2 scores for Sylheti (left) and Chittagonian (right) dialects under Zero Shot, Few Shot, and CoT prompting strategies.



(a) ROUGE-L Score comparison

(b) METEOR Score comparison

Figure 5: Performance comparison across different prompting strategies (Part 2). Comparison of ROUGE-L and METEOR scores for Sylheti (left) and Chittagonian (right) dialects under Zero Shot, Few Shot, and CoT prompting strategies.