Breaking the Silence: A Dataset and Benchmark for Bangla Text-to-Gloss Translation

Sharif Mohammad Abdullah[1], **Abhijit Paul**[1], **Shubhashis Roy Dipta**[2], **Zarif Masud**[3], **Shebuti Rayana**[4], **Ahmedul Kabir**[1]

[1]IIT, University of Dhaka, Bangladesh [2]University of Maryland, Baltimore County, USA
[3]Toronto Metropolitan University, Canada [4]SUNY, Old Westbury, USA

Correspondence: bsse1211@iit.du.ac.bd

Abstract

Gloss is a written approximation that bridges Sign Language (SL) and its corresponding spoken language. Despite a deaf and hard-of-hearing population of at least 3 million in Bangladesh, Bangla Sign Language (BdSL) remains largely understudied, with very limited prior work on Bangla text-to-gloss translation and no publicly available parallel dataset at the time of this work. To address this gap, we construct the first Bangla text-to-gloss dataset, consisting of 1,000 manually annotated and 4,000 synthetically generated Bangla sentence–gloss pairs, along with 159 expert human-annotated pairs used as a test set. Our experimental framework performs a comparative analysis between several fine-tuned open-source models and a leading closed-source LLM to evaluate their performance in low-resource BdSL translation. GPT-5.4 achieves the best automatic metric scores, while open source models perform competitively despite being much smaller. This work introduces the first publicly available expert-annotated Bangla text-to-gloss dataset, benchmark and trained model for Bangla text-to-gloss translation. It also demonstrates the effectiveness of systematically generated synthetic data for addressing challenges in low-resource sign language translation.[1]

1 Introduction

Sign language has been the primary method of communication for deaf people for centuries (Cantin and Encrevé, 2022). But deaf people often use lip reading, text-based communication (Barnett, 2002) to interact with others. Sign Language Generation (SLG), the process of converting text into sign, can improve the quality of life of deaf people by helping them receive information in their preferred language. In addition, the deaf community has expressed their preference in using sign over other media (Middleton et al., 2010).

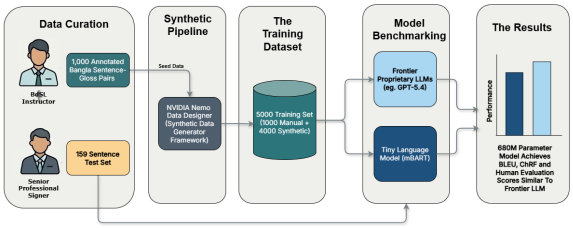


Figure 1: Our T2G pipeline: 1,000 expert-annotated pairs seed an LLM-based data generator to produce 4,000 synthetic sentences, for a total of 5,000 sentences. Fine-tuned models are evaluated against a proprietary LLM baseline using both automatic metrics and human judgment.

In the context of Sign Language (SL), gloss is considered as the written approximation of SL which uses words as "labels" for each sign along with various grammatical notes. A sign can have multiple meanings depending on the context of the sentence. For example, the sign RIGHT can mean a direction, a correct answer, or a legal/moral entitlement. "Your answer is correct." would be translated as "YOUR ANSWER RIGHT" and "Turn right at the light." would be translated to "TRAFFIC-LIGHT FORWARD TURN-RIGHT", even though the sign is same.

Bangla, a language spoken by 272.7 million individuals (Sultana et al., 2025), also has a substantial deaf community that uses the Bangla Sign Language daily. According to the government defined disability categories, among the people of Bangladesh, hearing disability is at

[1]https://github.com/SharifMAbdullah/breaking-the-silence

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 127–136
November 6-10, 2026 ©2026 Association for Computational Linguistics

0.29% (Bangladesh Bureau of Statistics), although Alauddin and Joarder (2004) suggest figures as high as 3 million. However, BdSL is a low-resource language. In our work, we introduce a novel dataset representing Bangla sentences and their gloss form. Egea-Gómez et al. (2021) shows that the use of glosses as intermediaries improves the translation process. To address this gap, we present following contributions,

- We curated a dataset of 1,000 hand-annotated Bangla sentences and their corresponding glosses, annotated by a certified BdSL instructor. To ensure a rigorous evaluation, a separate test set of 159 sentences was independently developed by a senior professional BdSL signer.
- We generate 4000 synthetic gloss with The NeMo Data Designer (2025) framework using the manually labeled data as the seed for the generation process
- We benchmark a diverse set of proprietary and open-source models to evaluate their performance on our dataset, and show that small models finetuned on our dataset can achieve performance comparable to proprietary LLM

Our dataset can serve as the foundation for the text-to-gloss-to-pose-to-video pipeline (Moryossef et al., 2023). Our results suggest that LLM-based synthetic data generation, seeded with a small expert-annotated dataset, can serve as a cost-effective complement to manual annotation for low-resource sign language tasks.

2 Literature Review

Text-to-Gloss. Text-to-gloss is a necessary task for sign language generation (SLG). Additionally, for tasks like text-to-sign, text-to-gloss translation is a necessary step. Moryossef et al. (2023) demonstrates that T2G is a crucial part in the text-to-gloss-to-pose-to-video pipeline. Despite such a necessity, text-to-gloss remains a low-resource domain. Egea-Gómez et al. (2021) propose a syntax-aware Transformer for English text-to-gloss translation, showing that incorporating syntactic information improves gloss generation quality. On the data side, Moryossef et al. (2021) address the scarcity of parallel gloss corpora by proposing rule-based heuristics that generate pseudo-parallel gloss-text pairs from monolingual spoken language text, demonstrating that pre-training on such synthetic data improves gloss translation by up to 3.14 BLEU. Our work builds on this line of research by applying LLM-based synthetic generation to BdSL, a language pair not previously studied.

Bangla Sign Language. Moryossef and Goldberg (2021) describes 20 distinct tasks for sign language translation. Among them, BdSL mainly focuses on the video (sign) to text task (Zeeon et al., 2024; Sams et al., 2023; Sonare et al., 2021; Alam et al., 2021; Hoque et al., 2016; Harini et al., 2020). Recent works include BdSLW60 prepared by Rubaiyeat et al. (2025a). This is an unconstrained word-level video dataset of 60 signs across over 9,000 trials. Building upon this work, Rubaiyeat et al. (2025c) scaled word recognition with BdSLW401 containing over 102,000 video samples. Another work by Rubaiyeat et al. (2025b) pioneered IsharaKhobor, a sentence translation dataset harvested from TV news. Few works have focused on the text-to-sign and text-to-pose tasks (Shahriar et al., 2017; Sarkar et al., 2009; Hoque et al., 2016). Islam et al. (2025) introduced IsharaKotha, the first-ever HamNoSys-based BdSL corpus of 3,823 words. However, Bangla lacks extensive research in gloss annotation, though recent work by Saha et al. (2026) introduced Bangla-SGP, a parallel corpus of sentence-gloss pairs built with rule-based pipeline. In fact, there is a general scarcity of text-to-gloss datasets across languages (Moryossef et al., 2021); we provide a comparison of existing corpora in App. B.

LLM for synthetic data generation. LLM-generated synthetic data has been shown to improve model performance when combined with human-labeled data, particularly under budget constraints (Matta et al., 2024; Zhezherau and Yanockin, 2024). However, smaller models tend to exhibit biases towards their own generated data, while larger models show more reliability (Maheshwari et al., 2024).

3 Methodology

3.1 Field Study

We started by interviewing BdSL teachers and professional signers. Key findings: (i) there is no single standard for BdSL – it varies between schools and individuals; (ii) BdSL follows rules of thumb rather than a strict grammar, such as using infinitive verb forms. Despite this, governmental and

academic institutions have published standardization guides. We curated part of our test set from these resources and the rest with a professional sign language interpreter, who re-verified every sentence. A separate annotator prepared the training data to reflect natural variation. We chose two annotators for the training and the test sets to prevent overfitting and to generalize the model. If the same person wrote both the training and test data, a model could achieve a good score by copying that person's style. But this would not have meant the model learned BdSL. Annotator details are in App. G.

3.2 Constructing Bangla T2G dataset

We construct a Bangla text-to-gloss (T2G) dataset which, to the best of our knowledge, is the first for Bangla language. We consider two data sources for Bangla T2G - human annotated data and synthetic data generated using LLM.

Curating hand annotated data

All contributors provided informed consent. We provided the training annotator with 1,000 Bangla sentences covering diverse aspects of everyday life and asked them to translate each into its gloss form. This manually created data served as ground truth and as seed for synthetic data generation.

Our final dataset comprises 1,000 human-annotated and 4,000 synthetic training pairs, with a separate test set of 159 human-annotated sentences (App. A).

Synthetic data generation using LLM

LLM has been widely used for generating synthetic data in low-resource domain and has shown sufficient reliability (Zehady et al., 2026; Alonso del Barrio et al., 2024). To find whether LLM can be used reliably to generate BdSL gloss, we take a sample of 100 sentences from the manually annotated dataset to generate gloss form for 100 Bangla sentences with The NeMo Data Designer (2025) framework and `GPT-4o` as the underlying LLM. In our prompt (provided in App. F), we mention the rules of thumb of BdSL, instruct the LLM to follow the patterns seen in the seed dataset and apply it all together to generate a new sentence-gloss pair for a new topic. We then evaluate these 100 text-to-gloss translations with our human annotator who created the training dataset who marked **75%** of sentences as correct. After gaining satisfactory remarks on this pilot dataset from our annotator, we generate gloss representations for an additional 4000 Bangla sentences using the same framework. The pilot sentences are not included in this dataset.

We combine the above two data sources to construct our final dataset (full details in Sec. 3.2), creating *bangla-gloss*.

3.3 Closed-source Baselines

`GPT-5.4` represents the state-of-the-art in generative LLM. It was incorporated into this study to serve as a state-of-the-art baseline for few-shot translation capabilities.

3.4 Trained Models for T2G Translation

We fine-tune four open-source models on our proposed training set: (i) **MarianMT** (Tiedemann and Nygaard, 2004), a dedicated NMT encoder-decoder pre-trained on OPUS; (ii) **mBART-50** (Liu et al., 2020), a multilingual seq2seq model whose bidirectional encoder and autoregressive decoder are well-suited for low-resource translation; (iii) **Qwen-3** (Yang et al., 2025), selected for its strong multilingual performance; and (iv) **Phi-4** (Abdin et al., 2024), included to assess whether a compact general-purpose LM can compete with dedicated translation models. Additional implementation details are provided in App. C.

3.5 Evaluation Scheme

Following Moryossef et al. (2021), we use three metrics - BLEU-4 (Papineni et al., 2002), chrF (Popović, 2015) and Human Evaluation to evaluate the text-to-gloss translation models. These metrics provide insight into the surface-level similarity of generated outputs with reference glosses, with higher-order BLEU scores emphasizing longer, contiguous n-gram matches. We do not use METEOR as their authors state that Bangla is not a supported language[2]. Although COMET supports Bangla, we do not use it because it hasn't seen BdSL in its training data, making it unsuitable for BdSL evaluation[3]. Full annotation and evaluation guidelines are provided in App. H.

4 Result & Analysis

The performance of all models on the *bangla-gloss* benchmark is summarized in Tab. 1. We organize our analysis around three key observations.

[2] https://www.cs.cmu.edu/~alavie/METEOR/
[3] github.com/Unbabel/COMET

Model	Type	Parameters	BLEU	chrF	Human Eval. (n=100)	95% CI
GPT-5.4	Causal LM	–	**39.26**	**73.75**	68%	[58.3, 76.3]
Qwen-3	Causal LM	8B	30.82	70.16	**72%**	**[62.5, 79.9]**
Phi-4 †	Causal LM	4B	2.26	26.95	0%	[0.0, 3.7]
mBART-large	Seq2Seq	0.68B	34.81	70.81	65%	[55.3, 73.6]
MarianMT†	Seq2Seq	0.14B	7.11	44.60	2%	[0.5, 7.0]

† Frequently produces misspelled tokens and incoherent gloss sequences, resulting in near-zero human acceptability.

Table 1: Automatic (BLEU-4, chrF) scores on the *bangla-gloss* test set (159 sentences) and human evaluation scores (100 sentences). Proprietary models are prompted few-shot; open-source models are fine-tuned on our training data. Human evaluation reflects acceptability as judged by a certified BdSL signer (App. G).

Automatic Metrics. GPT-5.4 achieves the highest BLEU (39.26) and chrF (73.75), establishing an upper bound among the models evaluated. Among open-source models, fine-tuned mBART attains the second-best BLEU (34.81) and chrF (70.81) despite being smaller, demonstrating the effectiveness of domain-specific fine-tuning of pretrained multilingual seq2seq models for low-resource translation. Qwen-3 follows closely with a BLEU of 30.82 and chrF of 70.16. MarianMT and Phi-4, by contrast, yield substantially lower scores, suggesting that neither a small dedicated NMT model nor a general-purpose small LM can adequately capture BdSL gloss structure under our training regime.

Human Evaluation. Human judgments reveal a notably different ranking. Qwen-3 receives the highest human evaluation score (72%), surpassing GPT-5.4 (68%) and mBART (65%). This suggests that Qwen-3 produces glosses that are semantically adequate to native signers even when their surface forms diverge from the reference. MarianMT and Phi-4 receive near-zero human scores (2% and 0%, respectively), as both models frequently generate misspelled tokens and incoherent output.

Metric Discordance. The gap between automatic and human scores, most notably Qwen-3 ranking first in human evaluation but third in BLEU, underscores a known limitation of reference-based metrics in gloss translation: because BdSL lacks a single standardized gloss convention, valid translations may differ lexically from the reference while remaining acceptable to native signers. Our overlap analysis (App. D) confirms the high lexical similarity between source text and gloss (mean Jaccard Similarity: 0.5487), which further limits the discriminative power of n-gram metrics. This finding aligns with prior observations on metric reliability in low-resource sign language processing (Moryossef et al., 2021) and motivates future work on reference-free evaluation for BdSL. Illustrative model translations are in App. E.

4.1 Ablation Study

To validate the contribution of each data source, we ablate on our two best-performing open-source models (Qwen-3 and mBART) under four configurations: pretrained (no fine-tuning), human-annotated only (1,000 sentences), synthetic only (1,000 subsampled), and the combined dataset (5,000 sentences). Results are in Tab. 2.

Model	Data Source	BLEU	chrF
Qwen-3	Pretrained	18.85	58.99
	Human	28.16	66.67
	Synthetic	29.78	68.31
	Combined	**30.82**	**70.16**
mBART-large	Pretrained	23.18	63.72
	Human	33.38	70.37
	Synthetic	29.78	67.26
	Combined	**34.81**	**70.81**

Table 2: Ablation on the two best-performing open-source models: pretrained, human-only, synthetic-only (1,000 each), and combined (5,000).

For both models, the combined dataset consistently yields the best performance. mBART improves from 23.18 (pretrained) to 33.38 (human-only) to 34.81 BLEU (combined). Qwen-3 shows a similar trend, from 18.85 to 30.82. Pretrained models perform substantially worse, confirming the necessity of domain-specific fine-tuning. Both human and synthetic data contribute complementary signal, as training on either alone yields lower scores than the combined dataset.

5 Conclusion & Future Work

We presented the first parallel dataset and benchmark for Bangla text-to-gloss translation, addressing a critical gap for a sign language commu-

nity of at least 3 million users. Our key finding is that a 680M-parameter mBART model, fine-tuned on just 4,000 sentences (80% of our 5,000-sentence training set), achieves performance competitive with GPT-5.4 on automatic metrics, while Qwen-3 surpasses all models in human evaluation. This demonstrates that domain-specific adaptation is the primary driver of quality in low-resource sign language generation. Our ablation study further shows that LLM-generated synthetic data provides complementary signal to expert annotations, supporting the viability of scalable data augmentation for sign language NLP. Future work will extend the dataset with regional BdSL variations and develop reference-free evaluation metrics for this domain.

Limitations

Gloss omits non-manual markers, so a gloss sequence accepted by a signer is a necessary but not sufficient condition for a fully faithful signed utterance. Although our dataset was annotated by people who know Bangla, it is possible that deaf people in other Bangla speaking area might follow different patterns of BdSL which are not represented in our work. Our dataset primarily consists of hand-annotated examples; however, the volume of such annotations remains limited. This constraint affects the coverage and diversity of linguistic constructions, particularly for complex or less frequently used sentence structures. Bangla exhibits unique syntactic and morphological characteristics that are reflected in Bangla Sign Language (BdSL). As such, a set of glossing rules tailored specifically to BdSL would likely produce more accurate and linguistically faithful representations. Due to the scarcity of qualified BdSL gloss annotators, we were unable to recruit a second independent evaluator for human assessment and therefore do not report inter-annotator agreement. Future work with an expanded annotator pool should address this gap.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Md Shahinur Alam, Mahib Tanvir, Dip Kumar Saha, and Sajal K Das. 2021. Two dimensional convolutional neural network approach for real-time bangla sign language characters recognition and translation. *SN Computer Science*, 2:1–13.

Mohammad Alauddin and Abul Hasnat Joarder. 2004. *Deafness in Bangladesh*, pages 64–69. Springer Japan, Tokyo.

David Alonso del Barrio, Max Tiel, and Daniel Gatica-Perez. 2024. Human interest or conflict? leveraging llms for automated framing analysis in tv shows. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, pages 157–167.

Bangladesh Bureau of Statistics. National survey on persons with disabilities (nspd) 2021. Accessed: 25-6-2025.

Steven Barnett. 2002. Communication with deaf and hard-of-hearing people: a guide for medical education. *Academic Medicine*, 77(7):694–700.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Yann Cantin and Florence Encrevé. 2022. Perspectives: On the historicalness of sign languages. *Frontiers in Communication*, 7:801862.

Santiago Egea-Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th workshop on building and using comparable corpora (BUCC 2021)*, pages 18–27.

R. Harini, R. Janani, S. Keerthana, S. Madhubala, and S. Venkatasubramanian. 2020. Sign language translation. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 883–886.

Md. Tazimul Hoque, Md. Rifat-Ut-Tauwab, Md. Fasihul Kabir, Farhana Sarker, Mohammad Nurul Huda, and Khandaker Abdullah-Al-Mamun. 2016. Automated bangla sign language translation system: Prospects, limitations and applications. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 856–862.

MD Islam, Prato Dewan, Md Fuadul Islam, Md Ataullha, and M Shahidur Rahman. 2025. Isharakotha: A comprehensive avatar-based bangla sign language corpus. *arXiv preprint arXiv:2511.16896*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *arXiv preprint arXiv:2409.11968*.

Shiho Matta, Yin Jou Huang, Fei Cheng, Hirokazu Kiyomaru, and Yugo Murawaki. 2024. Investigating cost-efficiency of llm-generated training data for conversational semantic frame analysis. *arXiv preprint arXiv:2410.06550*.

Anna Middleton, Graham H Turner, Maria Bitner-Glindzicz, Peter Lewis, Martin Richards, Angus Clarke, and Dafydd Stephens. 2010. Preferences for communication in clinic from deaf people: A cross-sectional study. *Journal of Evaluation in Clinical Practice*, 16(4):811–817.

Amit Moryossef and Yoav Goldberg. 2021. Sign Language Processing. https://sign-language-processing.github.io/.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023. An open-source gloss-based baseline for spoken to signed language translation. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*, pages 22–33.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*, pages 1–11.

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.

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.

Husne Ara Rubaiyeat, Hasan Mahmud, Ahsan Habib, and Md Kamrul Hasan. 2025a. Bdslw60: A word-level bangla sign language dataset. *Multimedia Tools and Applications*, 84(34):42399–42423.

Husne Ara Rubaiyeat, Hasan Mahmud, and Md Kamrul Hasan. 2025b. Bangla sign language translation: Dataset creation challenges, benchmarking and prospects. *arXiv preprint arXiv:2511.21533*.

Husne Ara Rubaiyeat, Njayou Youssouf, Md Kamrul Hasan, and Hasan Mahmud. 2025c. Bdslw401: Transformer-based word-level bangla sign language recognition using relative quantization encoding (rqe). *arXiv preprint arXiv:2503.02360*.

Neelavro Saha, Rafi Shahriyar, Nafis Ashraf Roudra, Saadman Sakib, and Annajiat Alim Rasel. 2026. Introducing a bangla sentence–gloss pair dataset for bangla sign language translation and research. 11(16):10457–10466.

Ataher Sams, Ahsan Habib Akash, and S. M. Mahbubur Rahman. 2023. Signbd-word: Video-based bangla word-level sign language and pose translation. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Biswajit Sarkar, Kaushik Datta, CD Datta, Debranjan Sarkar, Shashanka J Dutta, Indranil Das Roy, Amalesh Paul, Joshim Uddin Molla, and Anirban Paul. 2009. A translator for bangla text to sign language. In *2009 Annual IEEE India Conference*, pages 1–4. IEEE.

Rhythm Shahriar, AGM Zaman, Tanvir Ahmed, Saqib Mahtab Khan, and HM Maruf. 2017. A communication platform between bangla and sign language. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 1–4. IEEE.

Kritsada Singhapoo, Akarachai Inthanil, and Attapon Pillai. 2025. Fine-tuning ai models with limited resources. In *2025 11th International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, pages 148–151. IEEE.

Babita Sonare, Aditya Padgal, Yash Gaikwad, and Aniket Patil. 2021. Video-based sign language translation system using machine learning. In *2021 2nd International Conference for Emerging Technology (INCET)*, pages 1–4.

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*.

NVIDIA The NeMo Data Designer. 2025. Nemo data designer: A framework for generating synthetic data from scratch or based on your own seed data. https://github.com/NVIDIA-NeMo/DataDesigner. GitHub Repository.

Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus-parallel and free: http://logos. uio. no/opus. In *LREC*.

Ulrich von Agris and K. Kraiss. 2007. Towards a video corpus for signer-independent continuous sign language recognition. In *Gesture in Human-Computer Interaction and Simulation*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Iftekhar E Mahbub Zeeon, Mir Mahathir Mohammad, and Muhammad Abdullah Adnan. 2024. Btvsl: A novel sentence-level annotated dataset for bangla sign language translation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE.

Abdullah Khan Zehady, Shubhashis Roy Dipta, Naymul Islam, Safi Al Mamun, and Santu Karmaker. 2026. Banglallama: Llama for bangla language. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 73–89.

Alexey Zhezherau and Alexei Yanockin. 2024. Hybrid training approaches for llms: Leveraging real and synthetic data to enhance model performance in domain-specific applications. *arXiv preprint arXiv:2410.09168*.

A Dataset Breakdown

Dataset	Train	Test
Human Annotated	1,000	159
Synthetic	4,000	–

Table 3: Dataset split across gloss experiments.

B Existing SLT Corpora

Language Pair	Sentences	Signs
Signum DGS–German (von Agris and Kraiss, 2007)	780	450
RWTH-PHOENIX-Weather-2014T (Camgoz et al., 2018)	8,257	1,066

Table 4: Publicly available SLT corpora with gloss annotations and spoken language translations.

C Training Details

All models were fine-tuned with Low Rank Adaptation (LoRA) on the *bangla-gloss* training set. The models were trained on a single T4 GPU on Google Colaboratory. Due to resource constraints, we utilized the 4-bit quantized variant of the LLMs (specifically via Unsloth's optimized bitsAndBytes integration). This reduces memory overhead while preserving performance (Singhapoo et al., 2025). Model-specific details are provided below.

- **MarianMT (Helsinki-NLP/opus-mt-en-inc):** A transformer encoder-decoder architecture trained on OPUS, a multilingual corpus of openly available documents (Tiedemann and Nygaard, 2004). It serves as a dedicated NMT baseline, in contrast to the general-purpose LMs listed below.
- **mBART-50 (facebook/mbart-large-50):** We select mBART-50 for its bidirectional Transformer architecture. Its 12-layer encoder and 12-layer decoder (with 1024-dimensional hidden states) are explicitly optimized for sequence-to-sequence reconstruction, making it superior to models that rely on separately initialized decoders such as BERT (Liu et al., 2020).
- **Qwen-3 (unsloth/Qwen3-8B-unsloth-bnb-4bit):** Selected for its strong multilingual performance, particularly on Indo-European languages (Yang et al., 2025). The model was fine-tuned with a maximum sequence length of 2048.

Hyperparameter	Value
Batch size	8
Epochs	3
Optimizer	AdamW (8-bit)
Learning rate	2×10^{-5}
Train/validation split	80/20
Evaluation prompt	App. F
Checkpoint Used	Final checkpoint (last epoch)

Table 5: Details of hyperparameters used to train the models

- **Phi-4 (unsloth/Phi-4-mini-instruct-unsloth-bnb-4bit):** A compact general-purpose language model included to test whether small LMs can match dedicated translation models or larger LLMs on the T2G task.

The values of the hyperparameter used in finetuning the models:

D Sentence vs. Gloss Dissimilarity Metrics

To quantify the structural and lexical differences between the original spoken language sentences and their corresponding gloss versions, we employed two distinct metrics: Levenshtein Distance and Jaccard Similarity.

Mathematical Definitions

- Jaccard Similarity: Jaccard similarity is a statistical measure to quantify the similarity of two sets. If A and B represent the unique sets of tokens (words) in the sentence and the gloss, then Jaccard similarity,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

.

- Levenshtein Distance: The minimum number of single-character (or word-level) edits (insertions, deletions, or substitutions) required to change one string into another. For our study, we consider character-level distance.

Metric	Mean	Median
Jaccard Similarity	0.5487	–
Character Levenshtein Distance	5.4088	4.0000
Word Levenshtein Distance	1.8931	2.0000
Normalized Word Similarity	0.6271	0.6667

Table 6: Lexical and Structural Dissimilarity Metrics Between Sentences and Their Gloss Versions

From the data analysis, we observe:

1. **Shared Vocabulary with Systematic Divergence:** A mean Jaccard Similarity of 0.5487 and Normalized Word Similarity of 0.6271 show that BdSL gloss shares substantial vocabulary with spoken Bangla, but diverges systematically. This shared vocabulary is expected, as BdSL gloss uses Bangla lexemes as labels for signs.

2. **Morphological Transformations:** The mean Character Levenshtein Distance (5.41) is notably higher than the mean Word Levenshtein Distance (1.89), indicating that changes occur *within* words rather than through word-level reordering or substitution. This pattern is consistent with morphological flattening (e.g., stripping inflections from verbs like translating খাই into the root form খাওয়া), which is the primary operation in BdSL glossing.

E Illustrative Model Translations

The following examples showcase model's capacity to translate Bangla text to gloss.

Original: আমি বিস্কুট খাই
English: I eat biscuit
Reference: আমি বিস্কুট খাওয়া
Predicted: আমি বিস্কুট খাওয়া ✓

Original: তুমি কোথায় যাও?
English: Where do you go?
Reference: তুমি কোথায় যাও
Predicted: তুমি কোথায় যাও ✓

Original: আমি বৃষ্টিতে ছাতা নিয়ে বাইরে যাই
English: I go out in the rain with an umbrella
Reference: আমি বৃষ্টি ছাতা নিয়ে আসা বাইরে যাওয়া
Predicted: আমি বৃষ্টি ছাতা নেওয়া বাইরে যাওয়া ✓

F Prompt for BdSL Synthetic Gloss Generation

Prompt: Sentence-to-Gloss Mapping

You are a linguistically careful assistant trained to generate Bangla Sign Language (BdSL) gloss approximations from Bangla sentences. You are given a seed dataset with two columns:

{{Text}} translates to {{Gloss}}

Your task is to learn the mapping pattern between Sentence -> Gloss from the examples and then generate ONE new, high-quality gloss for ONE Bangla sentence for the topic {{Topics}}.

Strict instructions:
1. Use Bangla words, written in Bangla script.
2. Use the correct infinitive form of verbs wherever applicable. Example: "কাজ করে" -> "কাজ করা", "গাইতে" -> "গাওয়া". Do not invent verb forms not seen or implied in the seed data.
3. Follow BdSL-preferred ordering as observed in the seed Gloss column.
4. Do not paraphrase or add information.
5. Do not remove core meaning.
6. If a sentence is ambiguous, choose the most neutral, literal gloss.
7. No numbering, no quotes, no extra whitespace.
8. If the input sentence contains: multiple clauses, temporal expressions, conditionals, negation, modality, or questions, the gloss MUST preserve that complexity.
9. Final output MUST strictly follow this format:

BanglaSentence -> BanglaGloss

Constraints:
- Exactly ONE line.
- Exactly ONE "->".
- Left side MUST be the input sentence verbatim.
- Right side MUST be the generated BdSL gloss.
- A simpler gloss than the input sentence is INVALID.
- Any deviation is an error.

G Annotator Details

The training dataset was annotated by a graduate student from University of Dhaka's education research institute with professional experience at a government school for the deaf. The test set was developed in collaboration with a professional sign language interpreter currently active in Bangladesh National Television. While the annotators are hearing professionals, we mitigated 'hearing-centric' bias by validating our linguistic framework and findings through structured interviews with veteran BdSL educators. The primary annotator was recruited via institutional networks and was compensated at a rate exceeding local professional standards for specialized linguistic tasks, reflecting the high workload of training data preparation. The test set validator contributed their time on a voluntary basis as a professional consultant. All participants provided written informed consent.

They were briefed on the project's objectives, the data's intended use for academic research, and the anonymized nature of the eventual public release. Annotators were instructed to use the natural flow of BdSL. No personally identifiable information is included in the dataset. This study was conducted under the supervision of the authors' institutional department. In the absence of a formal Institutional Review Board (IRB) for non-medical linguistic research, the protocol followed the university's internal research code of conduct to ensure the protection of all participants.

H Evaluation Guidelines

Gloss annotation (dataset construction). Annotators were instructed to convert each Bangla sentence into the gloss form they would naturally use in their professional practice. The following conventions were followed:

1. Verbs are reduced to their infinitive/root form (e.g., খাই $\rightarrow$ খাওয়া).
2. Inflectional suffixes (case markers, tense markers) are stripped.
3. Punctuation marks are removed.
4. BdSL-preferred word ordering is maintained.
5. Function words are dropped where BdSL convention omits them.

These conventions were derived from the field study (Sec. 3.1) and the government-published guide "Isaray Bangla Bhasar Sahayika."

Human evaluation (model output assessment). The evaluator was presented with each model-generated gloss alongside the original Bangla source sentence. The reference gloss was *not* shown. For each output, the evaluator made a binary judgment: *acceptable* if the generated gloss matches how the evaluator would personally sign the sentence in BdSL, or *unacceptable* otherwise. The acceptability percentages reported in Tab. 1 reflect this judgment across all 100 test sentences presented to the evaluator. We did not present them with all 159 test sentences due to time and resource constraints.