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 | bsse1211@iit.du.ac.bd | AACL Main Track

Why Bangla text-to-gloss?

Gloss writes a sign language as word labels, one per sign. It is the entry step of **text** → **gloss** → **pose** → **video** (Moryossef et al., 2023).

Gloss is not word-for-word translation

The sign RIGHT has several meanings:
"Your answer is correct." → `YOUR ANSWER RIGHT`
"Turn right at the light." → `TRAFFIC-LIGHT FORWARD TURN-RIGHT`

BdSL is under-resourced:

- Hearing disability: **0.29%** of Bangladesh (official survey), estimates reach **up to 3 million** deaf people.
- Prior BdSL work is mostly video → text; no public expert-annotated text-to-gloss data.

Contributions

1. **First expert-annotated Bangla text-to-gloss dataset**: 1,000 hand-annotated + 4,000 synthetic pairs, and a 159-sentence expert test set.
2. **Benchmark**: four fine-tuned open models vs. few-shot GPT-5.4.
3. **Evidence** on metric reliability and on data sources (ablation).

Concurrent work: Bangla-SGP (Saha et al., 2026), a rule-based sentence–gloss corpus.

Field study: no single BdSL standard

Interviews with BdSL teachers and signers: BdSL **varies across schools and individuals** and follows rules of thumb. Our annotation conventions:

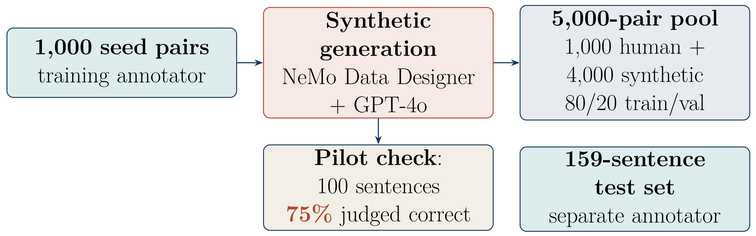
- Verbs reduced to infinitive/root form
- Inflectional suffixes stripped, punctuation removed
- BdSL-preferred word order; some function words dropped

আমি বিস্কুট খাই → আমি বিস্কুট খাওয়া
"I eat biscuit" verb → infinitive

Models and evaluation

- **Few-shot**: GPT-5.4
- **Fine-tuned** (LoRA, 4-bit, one T4 GPU, 3 epochs): mBART-50 (0.68B), Qwen3-8B, Phi-4-mini (4B), MarianMT (0.14B)
- **Metrics**: BLEU-4, chrF on 159 test sentences
- **Human eval**: The evaluator judged 100 model generated glosses

Building the dataset

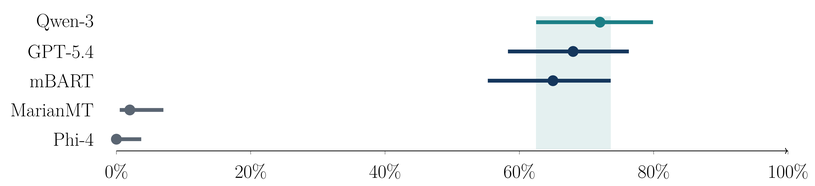


- **Training annotator**: Graduate student with experience at a government school for the deaf.
- **Test annotator**: Professional interpreter on national TV; part of the set from the government guide, every sentence re-verified.
- **Why two annotators for two dataset?**: On purpose. A shared annotator would let a model score well by copying one person's style.
- The validity of the synthetic pairs were verified by using a small subset(100 pairs)

Main results

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

† misspelled tokens, incoherent gloss. Human eval: $n = 100$.



Human-eval acceptability with 95% CIs; the top three overlap (shaded).

- mBART (0.68B) is **competitive** with few-shot GPT-5.4 on automatic metrics (34.81 vs. 39.26 BLEU).
- MarianMT(0.14B) and Phi-4(4B) performs poorly as they generate gibberish.

Metrics and humans disagree

Qwen-3 is **1st in human eval but 3rd in BLEU**: Valid glosses can differ from the single reference.

Prediction differs from reference; signer accepts it

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

- Sentence and gloss share many words (mean Jaccard **0.55**), which suggests limited discriminative power for n-gram metrics.
- Motivates **reference-free evaluation**.

Ablation: which data source helps?

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

- **Fine-tuning matters most**: about +9 to +10 BLEU over pretrained.
- Synthetic-only is close to human-only; combined is best by about 1.0 to 1.4 BLEU.
- Single run per cell; arms are 1,000 / 1,000 / 5,000 as reported, so combined's gain may partly reflect data volume.

Limitations and future work

Limitations: one judge, no inter-annotator agreement; human eval on 100 of 159 sentences; hearing annotators (mitigated by educator interviews); regional variation not covered; gloss omits non-manual markers.

Future work: regional BdSL, reference-free evaluation, linguist-informed glossing rules, equal-size data ablation.

Take-home and resources

Small expert seed + LLM-generated pairs gives fine-tuned models competitive with a frontier LLM on automatic metrics. **Without a gloss standard, BLEU/chrF mislead**: pair them with signer judgments.

`github.com/SharifMAbdullah/breaking-the-silence`

arXiv:2504.02293

LaTeX TikZposter