

Script Matters: Benchmarking Open-Source Machine Translation for Manipuri in Meetei-Mayek and Bengali Script

Kshetrimayum Boynao Singh^{1,2}, Asif Ekbal¹, Partha Pakray²

Department of Computer Science and Engineering

¹Indian Institute of Technology Patna, India

²National Institute of Technology Silchar, India

boynfrancis@gmail.com, asif@iitp.ac.in, partha@cse.nits.ac.in

Abstract

Manipuri (Meiteilon) officially recognised the Meetei-Mayek script alongside the Bengali script in 2021, with both scripts permitted for concurrent use during a 10-year transition period. However, the FLORES-200 devtest release still encodes Manipuri only as Bengali script, leaving the language’s current official script without a standard benchmark. We address this gap by contributing the (`mni_Mtei`) layer, aligned with the existing Bengali-script (`mni_Beng`) data, to form a four-way parallel resource comprising English, Hindi, Meetei-Mayek Manipuri, and Bengali-script Manipuri, with 1,012 sentences per language. This resource enables separate evaluation of open-source MT and LLM systems on the two Manipuri scripts, rather than treating them as a single “Manipuri” target. Across three IndicTrans2 checkpoint families, Sarvam-Translate, and five general-purpose LLMs, evaluated using six automatic metrics, we find that legacy Bengali-script output often scores higher than Meetei-Mayek despite the latter being the official script; IndicTrans2’s family degrades sharply on Bengali-script Manipuri; Sarvam-Translate cannot produce Bengali script at all; and none of the five general-purpose LLMs we test can reliably read or write Meetei-Mayek. Script identity is therefore a first-class evaluation variable for Manipuri MT: a single “Manipuri” score can mask large, direction-dependent quality gaps. We release our data¹, inference scripts, and evaluation harness.

1 Introduction

Manipuri (ISO 639-3: `mni`), spoken primarily in Manipur, India, is officially written in Meetei-Mayek (Unicode block `mni_Mtei`), reinstated as the state’s official script in 2021 after decades of Bengali-script (`mni_Beng`) usage following historical script suppression. FLORES-200 (Team

et al., 2022), the standard n -way parallel benchmark used to evaluate MT systems across low-resource languages, predates this switch: its Manipuri devtest split is encoded only as `mni_Beng`, so every existing FLORES-based Manipuri leaderboard number describes a script the language officially stopped using in 2021.

We fill this gap directly: we contribute a `mni_Mtei` devtest split aligned sentence-for-sentence to the original 1012-sentence FLORES-devtest split, alongside a matching Hindi split, so the same underlying content is available in English, Hindi, and *both* Manipuri scripts. This lets us evaluate current MT and LLM systems on Meetei-Mayek under the same FLORES-devtest protocol used for hundreds of other languages, and lets us hold everything except script constant when comparing `mni_Mtei` against `mni_Beng` output for the same model.

The stakes are not merely academic: a system deployed for Manipuri government or education services today is expected to produce Meetei-Mayek, since school and college curricula have used it as the medium of instruction since 2021, while a system trained on legacy web-crawled text may default to the wrong script for that use case. Meetei-Mayek is also, unlike Bengali script, used exclusively for Manipuri, which removes a specific source of ambiguity we return to in Section 4: several systems we test, when asked to produce Bengali-script Manipuri, instead default to producing plain Bengali or otherwise misidentify the language, an error that is structurally impossible when the target script is Meetei-Mayek. To our knowledge, no existing benchmark evaluates the same set of MT systems on both scripts side by side to quantify this gap.

We make the following contributions:

- We contribute, to our knowledge, the first publicly documented FLORES-devtest-

¹https://huggingface.co/datasets/helloboyn/flores-devtest-1012-manipuri-mni_Mtei

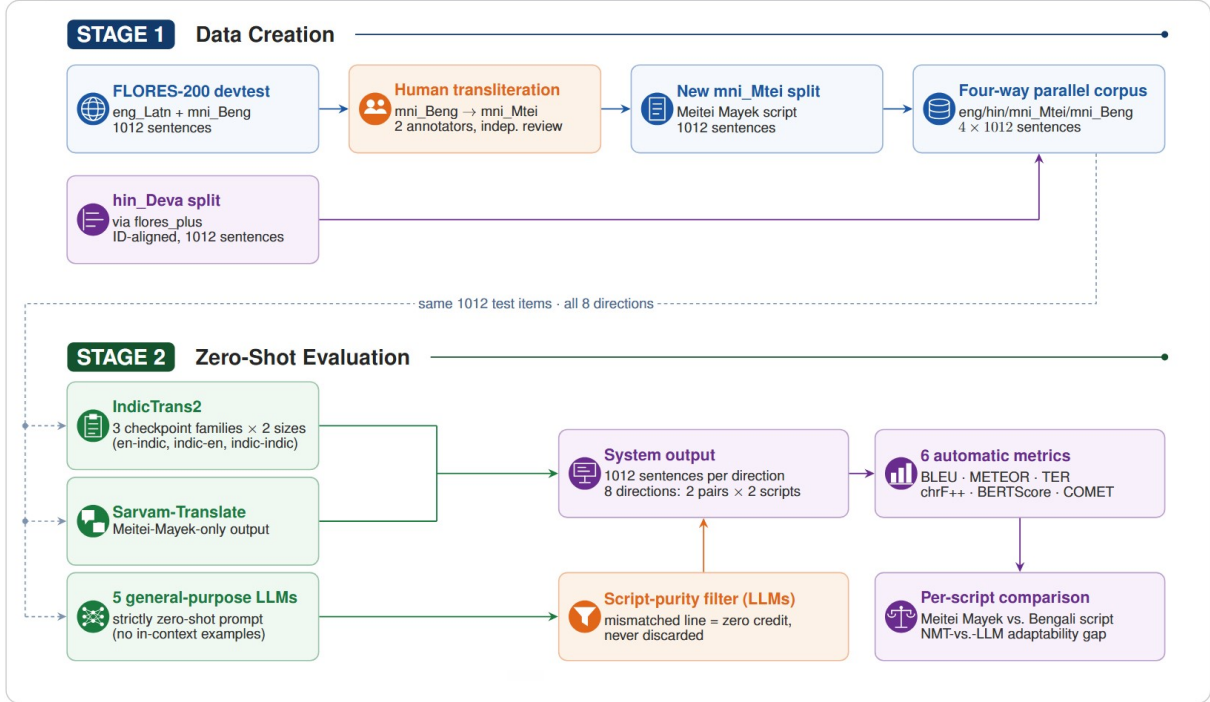


Figure 1: End-to-end pipeline. **Stage 1** extends FLORES-200 with an ID-aligned hin_Deva split and a human-transliterated mni_Mtei split, yielding a four-way parallel corpus (1012 sentences per language/script). **Stage 2** evaluates IndicTrans2, Sarvam-Translate and five LLMs on all eight directions; script-mismatched LLM lines are zero-credited rather than discarded, so every system is scored on the same 1012 sentences.

aligned Meetei-Mayek (mni_Mtei) evaluation split (human transliteration with independent second-annotator review, not independent re-translation), with a matching Hindi (hin_Deva) and English (eng_Latn) split, forming a four-way parallel resource (1012 sentences each). Wording is held fixed across the two Manipuri scripts by construction, so the resource also supports direct mni_Beng ↔ mni_Mtei transliteration evaluation.

- We benchmark IndicTrans2 (three checkpoint families, two sizes each) and Sarvam-Translate across both scripts and both translation directions, using six automatic metrics: BLEU, METEOR, TER, chrF++, BERTScore, and COMET.
- We benchmark five general-purpose instruction-tuned LLMs on the full 1012-sentence split across all eight directions with the same six metrics. None can reliably read or write Meetei-Mayek, and even the best LLM/direction result we score (chrF++ 25.76) falls short of the weakest reliable dedicated-MT result (chrF++ 32.89),

establishing this as currently a dedicated-MT-system task, not a general-LLM one.

- We document three qualitatively distinct failure modes that only appear on the Bengali-script side of dedicated MT systems: a checkpoint-specific hallucination artifact, register drift into Sanskritized vocabulary, and beam-search repetition collapse. None occur for the same models on Meetei-Mayek.
- We show that Sarvam-Translate, the strongest system on Meetei-Mayek, cannot be steered to produce Bengali-script Manipuri at all via prompting, and that dedicated MT and general-purpose LLMs favor *opposite* scripts on average – a training-data-curation effect, not an inherent property of either script.

2 Related Work

Multilingual MT benchmarks and Indic-focused MT. FLORES-200 (Team et al., 2022) provides n -way parallel evaluation data for 200 languages, including Manipuri, encoded as mni_Beng; its release predates Manipur’s 2021 reinstatement of Meetei-Mayek, so systems trained purely on FLORES/NLLB data (NLLB-200,

MADLAD-400 (Kudugunta et al., 2023), SeamlessM4T (Seamless Communication et al., 2023)) inherit Bengali-script-only support. IndicTrans2 (Gala et al., 2023), trained for India’s 22 scheduled languages, is to our knowledge the first widely used open MT system to treat Meetei-Mayek as a first-class target script, shipping separate checkpoint families for en-indic, indic-en, and indic-indic translation. Sarvam-Translate (Sarvam AI, 2025), a Gemma3-4B instruction-tuned model from Sarvam AI with AI4Bharat, lists Manipuri among its supported languages but per its model card supports only English↔Indic directly.

General-purpose LLMs for translation.

Instruction-tuned chat LLMs are increasingly used for translation without dedicated MT training, but published multilingual coverage claims rarely extend to recently-changed-script languages: none of the five LLMs we screen (Section 5) lists Meetei-Mayek as a supported language, so their behavior on mni_Mtei is, to our knowledge, previously undocumented.

Evaluation metrics. We report metrics with different sensitivities: surface n -gram overlap (BLEU; Papineni et al., 2002), edit distance (TER; Snover et al., 2006), synonym-aware overlap (METEOR; Banerjee and Lavie, 2005), character-level F -score (chrF++; Popović, 2017), and learned semantic similarity (BERTScore; Zhang et al., 2020, using XLM-R Conneau et al., 2020; COMET; Rei et al., 2020, 2022).

3 Data

Figure 1 gives an overview of the full pipeline, from data creation through evaluation.

We extend the FLORES-devtest split (1012 sentences) with four parallel, strictly line-aligned reference files: eng_Latn and mni_Beng from the original FLORES-200 devtest release; hin_Deva, sourced via openlanguageata/flores_plus and aligned by sentence ID; and mni_Mtei, likewise ID-aligned. This lets us evaluate any of the twelve ordered pairs among {English, Hindi, Manipuri-Meetei, Manipuri-Bengali} and, critically, isolate script as the only varying factor when comparing mni_Mtei against mni_Beng translations of the same content.

4 Script-Level Evaluation

mni_Mtei and mni_Beng are script-level counterparts of each other (the same underlying translation, transliterated rather than independently re-translated), unlike a prior 997-sentence Meetei-script overlap produced by independent translators for each script, where word choice can diverge and confound script identity with translator choice. The mni_Mtei split was produced by two native-speaker professional translators: one transliterated the FLORES-devtest mni_Beng split into Meetei-Mayek, and the second independently reviewed every sentence (Appendix A has the full transliteration guidelines).

5 Models and Experimental Setup

We evaluate two dedicated MT systems and five general-purpose LLMs. IndicTrans2 (Gala et al., 2023) ships three checkpoint families (en-indic, indic-en, indic-indic, each a 1B model plus a distilled 200–320M variant), with mni_Mtei and mni_Beng as distinct tokenizer language tags in every family. Sarvam-Translate (Sarvam AI, 2025) is prompted as a chat model and supports only English↔Indic directly; we pivot through English for Hindi↔Manipuri. The five general-purpose LLMs – Aya Expanse 8B (Cohere For AI, 2024), Qwen3-14B (Qwen Team, 2025), Gemma-3-12B-it (Gemma Team, 2025), Llama-3.1-8B-Instruct (Llama Team, AI @ Meta, 2024), and GLM-4-9B-0414 (Zhipu AI and Tsinghua KEG, 2024) – are evaluated with a role-primed prompt (Appendix D) across all eight directions (both scripts × both language pairs); none was screened out, since our goal is to characterize *how* they fail, not to find one that succeeds. Appendix B gives full checkpoint, decoding, and prompting configuration.

All inference uses greedy or low-temperature decoding, batch sizes of 8–16, and bfloat16 precision on a single GPU. We report six metrics per direction: BLEU and chrF++ (sacrebleu), TER (sacrebleu), corpus-averaged sentence METEOR (nltk), BERTScore F_1 (xlm-roberta-large), and COMET (wmt22-comet-da). The twenty dedicated-MT combinations are each run over the full 1012-sentence split, discarding and re-running any output with blank lines or a line-count mismatch; the five LLMs are run over the same split with a looser criterion, since a script-purity filter can legitimately zero out a failed line rather than

Direction	Model	chrF++↑	BLEU↑	METEOR↑	TER↓	BERTScore↑	COMET↑
eng → mni_Mtei	IndicTrans2-1B	33.16	3.37	16.12	112.12	97.64	85.24
	IndicTrans2-200M	32.89	3.43	15.83	109.68	97.60	84.74
	Sarvam-Translate	34.13	4.66	17.29	88.99	97.93	84.94
mni_Mtei → eng	IndicTrans2-1B	51.79	27.07	49.89	62.23	94.69	73.20
	IndicTrans2-200M	49.69	24.35	47.48	64.92	94.27	71.18
	Sarvam-Translate	42.62	18.33	39.29	70.25	93.40	69.35
hin → mni_Mtei	IndicTrans2-1B	30.42	2.55	13.90	121.48	97.23	85.22
	IndicTrans2-320M	29.68	2.35	13.43	122.34	97.24	85.57
	Sarvam-Translate	29.30	3.15	13.48	93.34	97.54	85.65
mni_Mtei → hin	IndicTrans2-1B	39.74	16.04	35.87	75.56	92.85	63.44
	IndicTrans2-320M	37.76	14.31	33.61	78.58	92.44	61.40
	Sarvam-Translate	34.28	12.17	29.84	78.79	92.04	61.62

Table 1: Automatic evaluation results on our proposed Manipuri Meetei-Mayek (mni_Mtei) dataset. BERTScore and COMET should be interpreted cautiously.

Direction	Model	chrF++↑	BLEU↑	METEOR↑	TER↓	BERTScore↑	COMET↑
eng → mni_Beng	IndicTrans2-1B	37.01	7.10	21.74	93.01	90.67	68.23
	IndicTrans2-200M	36.75	7.64	20.95	85.93	90.56	67.92
mni_Beng → eng	IndicTrans2-1B [†]	50.07	23.97	47.41	69.17	93.94	72.78
	IndicTrans2-200M [†]	47.06	21.09	43.79	72.05	93.40	70.01
hin → mni_Beng	IndicTrans2-1B [‡]	9.85	0.42	0.32	109.70	82.64	46.82
	IndicTrans2-320M [‡]	4.10	0.30	0.49	136.76	80.03	31.46
mni_Beng → hin	IndicTrans2-1B [‡]	36.09	11.51	32.61	93.73	91.37	57.88
	IndicTrans2-320M [‡]	32.66	8.76	29.56	103.99	90.06	51.66

Table 2: Translation performance on the Bengali-script variant for English–Manipuri and Hindi–Manipuri directions; Sarvam-Translate is excluded due to lack of support, and results are reported for models supporting the Bengali-script variant.

signal an interrupted run – Appendix B details both criteria.

Our central finding is that dedicated MT and general-purpose LLMs favor *opposite* scripts on average. Full per-model, per-direction tables (twelve Meetei-Mayek and twenty Bengali-script/LLM directions) are in Appendix C.

The legacy script scores higher on English→Manipuri. Both IndicTrans2 sizes score eng→mni_Beng above eng→mni_Mtei (BLEU 7.10–7.64 vs. 3.37–3.43), even though Meetei-Mayek is the current official script – we attribute this to training-data recency asymmetry (Appendix C). Notably, BERTScore and COMET *disagree* with this ordering here, which we read as encoder unfamiliarity with both scripts at this quality range rather than a genuine vote for Meetei-Mayek.

Two Bengali-script-only failure modes in IndicTrans2. The indic-en checkpoint intermittently hallucinates a literal escape-like fragment (e.g., ` ` at Unicode code point U+09BC) into English output when the mni_Beng source contains a foreign proper noun – confirmed as a decoding-time artifact, not present in the source, and absent on mni_Mtei input. The indic-indic family (direct Hindi↔Manipuri) shows our largest script-dependent gap: hin→mni_Beng collapses to near-zero BLEU (0.30–0.42) via Sanskritized register drift and, for the smaller checkpoint, script-mixing and repetition loops, versus 2.35–2.55 BLEU on hin→mni_Mtei. Appendix C gives full examples of both artifacts.

Sarvam-Translate cannot produce Bengali script at all. Across all four Bengali-script directions, explicitly instructing the model to translate to or from “Manipuri (Bengali script)”

Direction	Model	chrF++↑	BLEU↑	METEOR↑	TER↓	BERTScore↑	COMET↑
eng → mni_Mtei	Aya-Expanse-8B	0.60	0.08	0.03	108.35	77.80	20.99
	Gemma-3-12B-it	3.56	0.00	0.25	100.91	38.45	40.83
	Llama-3.1-8B-Instruct	0.34	0.08	0.08	102.85	77.55	19.03
	Qwen3-14B	0.29	0.05	0.04	112.20	89.28	52.25
	GLM-4-9B-0414	0.36	0.00	0.03	102.09	4.12	17.59
mni_Mtei → eng	Aya-Expanse-8B	17.70	0.18	7.89	216.76	84.81	42.10
	Gemma-3-12B-it	19.95	0.65	10.24	133.51	87.02	47.13
	Llama-3.1-8B-Instruct	17.37	0.27	10.16	279.74	84.73	43.71
	Qwen3-14B	15.53	0.21	9.64	320.89	85.18	44.64
	GLM-4-9B-0414	19.77	0.62	10.81	157.38	86.50	45.50
hin → mni_Mtei	Aya-Expanse-8B	0.64	0.11	0.04	112.28	78.04	19.64
	Gemma-3-12B-it	4.53	0.02	0.28	103.09	49.49	50.48
	Llama-3.1-8B-Instruct	0.42	0.06	0.03	103.20	78.10	18.13
	Qwen3-14B	0.39	0.05	0.02	158.08	89.38	56.57
	GLM-4-9B-0414	0.08	0.00	0.01	100.55	1.14	15.96
mni_Mtei → hin	Aya-Expanse-8B	13.58	0.24	5.90	154.28	85.39	40.47
	Gemma-3-12B-it	6.73	0.30	3.38	120.47	84.40	36.89
	Llama-3.1-8B-Instruct	14.65	0.28	6.35	150.81	84.89	40.76
	Qwen3-14B	13.46	0.11	4.69	128.28	84.37	35.57
	GLM-4-9B-0414	14.41	0.37	6.25	121.29	84.89	35.57

Table 3: Zero-shot LLM screening on Meetei-Mayek translation directions using the full 1,012-sentence devtest set. As in Table 1, BERTScore/COMET should be read cautiously for mni_Mtei rows.

still produces fluent Meetei-Mayek; we treat this as a capability gap and report no Bengali-script Sarvam-Translate numbers.

General-purpose LLMs cannot read or write Meetei-Mayek. A script-purity filter (emptying lines where fewer than half the letters match the target script) scores 0.00–0.01 purity on the overwhelming majority of mni_Mtei-target lines across all five LLMs – total failure, not marginal. On generation, models mix in characters from unrelated scripts (Odia, Devanagari, Thai, Korean, and others); chrF++ never exceeds 4.53. On comprehension, models instead *confabulate* fluent, grammatical, entirely unrelated English/Hindi text rather than declining – e.g., a source about a diabetic mouse study translated as a fabricated paragraph about a village in Assam. Bengali-script comprehension is comparatively less broken (every model scores higher on mni_Beng→eng/hin than the mni_Mtei equivalent) but still unreliable for translation. Appendix C has the full per-model metric tables.

6 Discussion

Two lessons follow. First, *script cannot be assumed constant across a language’s benchmark entries*: a “Manipuri” score, reported without qual-

ification, conflates two eras of training data and, as our indic-indic results show, potentially two very different quality regimes for the same model. Second, *failure modes are script-specific, not just magnitude differences*: the hallucination, register-drift, and repetition-collapse artifacts we observe on Bengali-script input have no Meetei-Mayek counterpart for the same checkpoints, suggesting data scarcity interacting with architecture rather than uniform degradation.

Which script is more adaptable depends on the system class. IndicTrans2 favors Meetei-Mayek in three of four direction buckets (Sarvam-Translate, being Meetei-only, pushes the dedicated-MT picture further that way), while all five LLMs favor Bengali script in all four, with roughly triple the gap on generation versus comprehension. We read this as evidence that script adaptability reflects *training-data curation*, not either script’s inherent suitability: AI4Bharat built mni_Mtei into IndicTrans2 as a first-class target, while the LLMs’ pretraining absorbed whatever legacy Bengali-script Manipuri text was already circulating online. Practitioners should expect a dedicated MT system and a general-purpose LLM to favor opposite scripts by default (full numeric breakdown in Appendix C). We suggest that benchmarks

Direction	Model	chrF++↑	BLEU↑	METEOR↑	TER↓	BERTScore↑	COMET↑
eng → mni_Beng	Aya-Expans-8B	9.98	0.07	0.33	108.47	81.32	38.95
	Gemma-3-12B-it	8.14	0.12	0.36	106.31	81.40	42.98
	Llama-3.1-8B-Instruct	7.92	0.09	0.20	103.02	80.96	40.51
	Qwen3-14B	8.38	0.09	0.23	107.91	73.56	39.84
	GLM-4-9B-0414	6.69	0.02	0.17	102.80	50.93	36.31
mni_Beng → eng	Aya-Expans-8B	19.02	1.38	6.59	112.18	83.81	42.75
	Gemma-3-12B-it	25.76	3.09	13.87	114.65	88.05	54.35
	Llama-3.1-8B-Instruct	16.65	1.03	4.02	108.31	80.65	34.90
	Qwen3-14B	21.18	2.43	8.63	106.05	83.08	41.81
	GLM-4-9B-0414	21.22	1.87	9.48	119.94	85.32	45.70
hin → mni_Beng	Aya-Expans-8B	10.67	0.12	0.38	112.58	80.55	38.48
	Gemma-3-12B-it	9.68	0.15	0.45	109.05	81.36	42.43
	Llama-3.1-8B-Instruct	8.18	0.07	0.19	104.68	80.06	39.62
	Qwen3-14B	8.86	0.10	0.17	113.62	80.88	39.60
	GLM-4-9B-0414	6.91	0.06	0.16	105.60	54.89	36.29
mni_Beng → hin	Aya-Expans-8B	14.93	0.73	4.75	103.44	83.92	35.79
	Gemma-3-12B-it	17.59	1.83	7.61	101.98	84.56	38.88
	Llama-3.1-8B-Instruct	13.77	0.58	2.52	100.19	81.87	31.25
	Qwen3-14B	13.76	0.22	3.94	104.84	82.90	32.55
	GLM-4-9B-0414	13.58	0.52	3.39	101.90	82.73	31.61

Table 4: Zero-shot evaluation of five general-purpose instruction-tuned LLMs on Bengali-script Manipuri across four translation directions. Results are reported on the full 1,012-sentence FLORES-devtest set using chrF++, BLEU, METEOR, TER, BERTScore, and COMET, highlighting the models’ ability to generate and comprehend Manipuri written in Bengali script.

for languages that have undergone a script transition report script explicitly as an evaluation axis alongside language pair, analogous to domain or genre in MT evaluation. As future work, we plan to extend the same script-paired methodology to the NLLB Seed/OLDI (Open Language Data Initiative) datasets, bringing Meetei-Mayek coverage to these broader low-resource-focused resources as well.

7 Conclusion

We contribute the first publicly documented FLORES-devtest-aligned Meetei-Mayek split, produced by human transliteration and independent review, paired with a matching Hindi split for direct, script-controlled evaluation of Manipuri MT. Benchmarking IndicTrans2, Sarvam-Translate, and five general-purpose LLMs across six automatic metrics, we show that script identity substantially changes both quantitative scores and qualitative failure modes for the same models: the legacy Bengali-script convention sometimes outperforms the current official script on surface metrics, the indic-indic checkpoint family and every LLM we tested

collapse specifically on Bengali-script Manipuri or Meetei-Mayek respectively, and dedicated MT and general-purpose LLMs favor opposite scripts on average. A single “Manipuri” number, as reported by prior benchmarks built on mni_Beng alone, would have missed all of this. We release the dataset, inference scripts, and evaluation harness so script can be evaluated as a first-class variable for Manipuri, and, as future work, for other languages that have undergone a similar script transition.

Limitations

BERTScore and COMET rely on multilingual encoders (XLM-R, and the COMET encoder) with, to our knowledge, minimal-to-no training exposure to Meetei-Mayek script, which was not in wide digital circulation until after these encoders’ pretraining corpora were collected. We report these metrics for completeness but recommend treating their absolute values on mni_Mtei rows as low-confidence; chrF++ and BLEU are the more trustworthy metrics for Meetei-Mayek here. Bengali-script rows are on firmer footing since Bengali is well represented in both encoders.

Our study covers a fixed set of open-source systems available at the time of writing and a single, general-domain test set (FLORES-devtest); results may not transfer to domain-specific or dialectal Manipuri text.

Acknowledgments

We thank the NLLB/FLORES-200 team for the original 1012-sentence devtest corpus and its Bengali-script Manipuri translations, on top of which we contribute the Meetei-Mayek layer that is the subject of this paper; the Open Language Data Initiative (OLDI) for accepting our proposal to contribute this resource; the COIL-D project for providing resources used in this work; and the AI4Bharat and Sarvam AI teams for releasing open, gated-but-accessible model weights that made this comparison possible.

References

- 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 Machine Translation and/or Summarization*, pages 65–72.
- Cohere For AI. 2024. **Aya expande: Combining research breakthroughs for a new multilingual frontier**. *arXiv preprint arXiv:2412.04261*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. **Unsupervised cross-lingual representation learning at scale**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. **Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages**. *Transactions on Machine Learning Research*.
- Gemma Team. 2025. **Gemma 3 technical report**. *arXiv preprint arXiv:2503.19786*.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. **MADLAD-400: A multilingual and document-level large audited dataset**. *arXiv preprint arXiv:2309.04662*.
- Llama Team, AI @ Meta. 2024. **The llama 3 herd of models**. *arXiv preprint arXiv:2407.21783*.
- 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. 2017. **chrF++: words helping character n-grams**. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.
- Qwen Team. 2025. **Qwen3 technical report**. *arXiv preprint arXiv:2505.09388*.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. **COMET-22: Unbabel-IST 2022 submission for the metrics shared task**. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.
- 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.
- Sarvam AI. 2025. **Sarvam-translate: An open indic translation model**. <https://huggingface.co/sarvamai/sarvam-translate>. Model card.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, et al. 2023. **SeamlessM4T: Massively multilingual & multimodal machine translation**. *arXiv preprint arXiv:2308.11596*.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. **A study of translation edit rate with targeted human annotation**. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. **No language left behind: Scaling human-centered machine translation**.

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*.

Zhipu AI and Tsinghua KEG. 2024. [ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools](#). *arXiv preprint arXiv:2406.12793*.

A Transliteration Guidelines

A.1 Transliteration and Review (mni_Mtei Split Creation)

Given to both annotators (Section 4), each a native-speaker professional translator holding a graduate degree and at least five years of translation experience:

1. **Task:** Transliterate each sentence in the FLORES-devtest mni_Beng (Bengali-script Manipuri) split into Meetei-Mayek script. This is a script conversion, not a re-translation: preserve the exact wording, word order, and meaning of the mni_Beng source sentence. Do not paraphrase, simplify, or "improve" the phrasing.
2. **Numerals and Latin text:** Retain digits, proper nouns in Latin script, and punctuation exactly as they appear in the source unless Meetei-Mayek orthographic convention requires a native numeral or punctuation mark; when in doubt, prefer the form already used in existing Meetei-Mayek educational text (see the reference textbook list provided separately to annotators).
3. **Ambiguous or archaic mni_Beng spellings:** Where the Bengali-script source is ambiguous between two Meetei-Mayek renderings, choose the rendering consistent with current Manipur school-curriculum orthography; flag the sentence ID in a shared tracking sheet for second-annotator discussion rather than resolving it unilaterally.
4. **Review pass (second annotator):** Independently read every transliterated sentence against the mni_Beng source (not against the first annotator's output in isolation) and mark each as Accept, Minor Edit (typo, incorrect character, orthography), or Flag for Discussion (meaning-level disagreement). Apply Minor Edits directly; resolve Flagged items jointly with the first annotator before finalizing.

5. **Output:** One Meetei-Mayek sentence per line, line-aligned to the source mni_Beng file, with no added commentary, footnotes, or alternate readings in the released file.

B Model and Screening Details

B.1 IndicTrans2 checkpoints

IndicTrans2 ships three separate checkpoint families, each independently gated on HuggingFace, because AI4Bharat trains distinct weights per translation direction rather than a single shared encoder-decoder: **en-indic** (English→Indic, 1B and a distilled 200M variant), **indic-en** (Indic→English, 1B and 200M), and **indic-indic** (Indic→Indic, used here for Hindi↔Manipuri, 1B and a distilled 320M variant). We verified directly against the tokenizer vocabulary that both mni_Mtei and mni_Beng exist as distinct language tags in every family, confirming Bengali-script support is a genuine modeling target rather than an accidental byte-fallback artifact. Inference uses the IndicTransToolkit pre/post-processing pipeline (normalization, script-aware digit conversion).

B.2 Sarvam-Translate prompting

Sarvam-Translate is prompted as a chat model ("Translate the text below to {language}."). with greedy-leaning decoding (temperature=0.01). For Hindi↔Manipuri we implement the model card's recommended two-hop English pivot: two forward passes per sentence with the same checkpoint.

B.3 Completeness criteria

Each of the twenty dedicated-MT (model, checkpoint, direction, script) combinations is run over the full 1012-sentence devtest split; we discard and re-run any output containing blank lines or a line-count mismatch against the reference, treating this as an interrupted run rather than a model failure. The five LLMs are run over the same full split across all eight directions each (forty (model, direction) combinations total), scored with the same six metrics; a purity-filtered (emptied) line is scored as zero-credit against the reference rather than discarded, so a model that fails on most sentences still receives a real, near-floor score. Here, only a line-count *mismatch* against the reference – not the presence of blanks – causes us to exclude a combination, since a blank can be a legitimate scored failure rather than an interrupted run. All forty LLM

(model, direction) combinations pass this check and are reported in Appendix C’s tables.

C Extended Results and Failure-Mode Analysis

This appendix expands the summarized findings of Section 4 with full per-model, per-direction metric tables and the detailed evidence behind each failure mode.

C.1 English→Manipuri: legacy script vs. official script

Both IndicTrans2 sizes score eng→mni_Beng noticeably above eng→mni_Mtei (BLEU 7.10–7.64 vs. 3.37–3.43; chrF++ 36.75–37.01 vs. 32.89–33.16), even though Meetei-Mayek is the script IndicTrans2 was explicitly built to target and the one with current official status. We attribute this to training-data recency asymmetry: Bengali-script Manipuri text has circulated on the web for decades and is well represented in legacy corpora, while Meetei-Mayek web text is comparatively young (large-scale digital adoption began only after 2021). The reverse direction (Manipuri→English) shows the same ordering but with a smaller gap (BLEU 21.09–23.97 for Bengali script vs. 24.35–27.07 for Meetei-Mayek), suggesting source-side script recognition is more robust than target-side script generation for the less-represented script.

Notably, the learned metrics *disagree* with this ordering on eng→mni: both BERTScore and COMET rank Meetei-Mayek output above Bengali-script output (COMET 84.7–85.2 vs. 67.9–68.2), the opposite of BLEU and chrF++. Since absolute quality is low on both sides (chrF++ in the low-to-mid thirties throughout), we read this as evidence that the multilingual encoders behind BERTScore and COMET are not well calibrated to discriminate translation quality in either script at this quality range, rather than a genuine vote for Meetei-Mayek.

Averaging across both IndicTrans2-1B sizes makes the asymmetry precise: on eng→mni, Bengali script leads by +3.73 BLEU and +3.85 chrF++ but trails by −17.01 COMET; on mni→eng, Meetei-Mayek leads on every metric, but by a much smaller margin (+3.10 BLEU, +1.72 chrF++, +0.42 COMET). The size of the COMET gap on eng→mni (16.91 points) dwarfs the corresponding gap on mni→eng (0.42 points)

even though the lexical-overlap metrics move by comparable amounts in both directions – a further sign that COMET’s judgment on eng→mni is dominated by encoder unfamiliarity with the target-side script, rather than by anything in the source.

C.2 A checkpoint-specific hallucination on Bengali-script input

The indic-en checkpoint (Manipuri→English, both sizes), when given mni_Beng input containing a foreign proper noun (person or company name), intermittently emits a literal escape-like fragment resembling spliced into the English output. We confirmed via direct inspection that ঐ (Bengali sign nukta) does not appear anywhere in the corresponding source sentence, ruling out a source-data artifact; this is a decoding-time hallucination specific to this checkpoint on this script, and the identical model on mni_Mtei input never exhibits it. We mitigate it with a post-hoc regular-expression cleanup (stripping the fragment and any stray backslash) in the released inference scripts, which raises usable output rate substantially, though sentence content lost during the original hallucination is not recoverable by this cleanup. Rows marked † in Table 2 reflect this mitigation.

C.3 indic-indic degrades sharply on Bengali-script Manipuri

The indic-indic checkpoint family, used for direct Hindi↔Manipuri translation, shows the largest script-dependent gap we observe. On hin→mni_Beng, both sizes collapse to near-zero BLEU (0.30–0.42) and very low chrF++ (4.10–9.85), an order of magnitude below the same models’ hin→mni_Mtei scores (2.35–2.55 BLEU). Manual inspection shows the 1B checkpoint drifting into a Sanskritized register with Sanskrit function words and verb morphology rather than producing colloquial Manipuri, while the 320M checkpoint additionally mixes Bengali, Meetei-Mayek, and Latin scripts within a single sentence and, on at least one output, degenerates into a repeated-token loop (the same Hindi word repeated over fifty times). The reverse direction (mni_Beng→hin) is degraded but not catastrophic (BLEU 8.76–11.51 vs. 14.31–16.04 for mni_Mtei→hin), consistent with the source-recognition-is-more-robust-than-target-generation pattern noted above, compounded by numeral corruption reminiscent of the artifact in

Appendix C.2 manifesting in Hindi rather than English.

C.4 General-purpose LLMs cannot read or write Meetei-Mayek

We probed five general-purpose instruction-tuned chat LLMs (Aya Expanse 8B, Qwen3-14B, Gemma3-12B, Llama-3.1-8B-Instruct, and GLM-4-9B) on the full 1012-sentence FLORES-devtest split across all eight directions each, using an explicit, role-primed system prompt (Appendix D). To separate genuine script failures from formatting noise, we built a script-purity filter that classifies each output character by Unicode block and empties any line where fewer than half of its letters belong to the target script or a recognizable Latin romanization; every model we tested this way scored a purity of 0.00–0.01 on the overwhelming majority of `mni_Mtei`-target lines, confirming the failure is total rather than marginal. A purity-filtered (emptied) line is scored as zero-credit against the reference rather than discarded, so a model that fails on most sentences still receives a real, near-floor corpus score rather than an inflated average over its rare successes.

Two distinct failure modes emerged, one on each side of the pair. On *generation* (English/Hindi→Manipuri), all five models mix characters from several unrelated scripts within a single line: we observed fragments of Odia, Bengali, Devanagari, Gujarati, Punjabi, Lao, Thai, Korean, Ol Chiki, and Thaana in outputs nominally targeting Meetei-Mayek, none of which is the requested script. Table 3 makes this quantitative: `chrF++` on `eng/hin→mni_Mtei` never exceeds 4.53, an order of magnitude below the weakest dedicated MT system in Table 1. On *comprehension* (Manipuri→English/Hindi), the failure is more concerning: rather than producing garbled output, models generate fluent, grammatical English or Hindi that is *entirely unrelated to the source sentence*. For example, one model’s translation of a source about a diabetic mouse study returned a fabricated paragraph about a village in Assam’s Naga hills, and another invented an unrelated news item about a WHO pandemic-preparedness designation. This is confident confabulation, not a failed translation attempt: the models cannot parse Meetei-Mayek as input at all, and default to generating plausible-sounding but unrelated text rather than declining or hedging; the resulting `mni_Mtei→eng/hin` `chrF++`

scores (6.73–19.95) are not zero, but reflect fluent off-topic text matching stray reference words by chance, not usable translation. Bengali-script Manipuri comprehension is comparatively less broken: several models produce recognizable phonetic transliterations of the source or retain correct proper nouns, and Table 4 shows every model scoring higher on `mni_Beng→eng/hin` than on the `mni_Mtei` equivalent, though absolute quality remains too low for usable translation. We conclude that Meetei-Mayek support in current general-purpose LLMs is not a prompting problem but a training-data coverage gap, consistent with the script’s short digital history, and that dedicated MT systems remain necessary for this language pair.

C.5 Which script is more adaptable: full numeric breakdown

Averaging `chrF++` within each direction bucket across all evaluated models reveals a split between dedicated MT systems and general-purpose LLMs. For IndicTrans2, Meetei-Mayek is the better-supported script in three of four buckets: `hin→mni` favors Meetei-Mayek by +23.1 `chrF++` (the *indic-indic* collapse from Appendix C.3), `mni→hin` by +4.4, and `mni→eng` by +2.2; only `eng→mni` favors Bengali script, by +3.9. Sarvam-Translate is Meetei-only by construction, pushing the dedicated-MT picture further toward Meetei-Mayek. The five general-purpose LLMs show the opposite and more consistent pattern: Bengali script wins all four buckets, by +7.2 `chrF++` on `eng→mni`, +7.6 on `hin→mni`, +2.7 on `mni→eng`, and +2.2 on `mni→hin`—generation shows roughly triple the comprehension-direction gap, the same source-recognition-is-more-robust-than-target-generation asymmetry noted above, just uniformly shifted toward the legacy script.

D Prompt Templates and Zero-Shot Examples

For the general-purpose LLM screening in Appendix C.4, each of the eight direction-specific prompts follows the same role-primed template, instantiated with the appropriate source/target language names:

“Translate the text below to {target language} (target language). Output only the translation, with no explanation or extra text.”