

Script Is Not Preprocessing: Tokenizer-Mediated Transfer in Endangered Language Machine Translation

Pooja Singh¹ Saurabh Chauhan¹ Nishant Agnihotri¹ Sandeep Kumar^{1,3}

¹Indian Institute of Technology Delhi, New Delhi, India

³PiByte AI, New Delhi, India

eez228470@ee.iitd.ac.in saurabhchauhan1089@gmail.com

nishantagnihotri38@gmail.com ksandeep@iitd.ac.in

Abstract

Multilingual machine translation for endangered and severely underrepresented languages faces two interacting challenges: scarce parallel data and uneven subword vocabulary allocation across writing systems. We study these factors for seven Indian tribal languages, Gondi, Mundari, Bhili, Santali, Kui, Garo, and Kokborok, spanning four language families and four scripts. We release a speaker-curated Hindi-English-X tri-parallel corpus of approximately 20k sentences per language across six community-relevant domains, with all development and test references human-curated or independently speaker-verified. We evaluate prompted LLMs, fine-tuned IndicTrans2 and NLLB-600M, and byte-level ByT5 under corpus-Native, Devanagari-Unified, and Latin-Unified script conditions. Fine-tuning improves chrF++ by 16-25 points over prompting, but script unification is strongly model-dependent: the same Latin conversion improves Kui under NLLB-600M while substantially degrading performance under IndicTrans2. We introduce Script Tokenization Parity and Script Sensitivity to analyse script-tokenizer interaction; byte-level modelling reduces Script Sensitivity from 14.2 to 2.1 chrF++, localizing most cross-script variance in learned vocabulary allocation. Our findings show that script choice in endangered-language MT is not routine preprocessing, but a community-sensitive modelling decision that directly shapes linguistic representation, translation quality, and equitable digital access.

1 Introduction

India’s census records more than 460 tribal languages spoken by over 100 million people ¹, yet

¹Census of India 2011, C-16: Population by Mother Tongue, Office of the Registrar General & Census Commissioner, India: <https://censusindia.gov.in/nada/index.php/catalog/10191>.

nearly all lack parallel corpora, standardised evaluation suites, and meaningful representation in the pretraining distributions of multilingual systems (Khanuja et al., 2021, 2023; Khan et al., 2024). The consequences of this absence are immediate for digital access. In practice, it affects whether agricultural advisories, public-health protocols, governance information, and language-documentation materials can be produced and accessed directly in the relevant community language, rather than mediated through a high-resource language such as Hindi or English. This gap limits access to education, health information, governance, and digital participation, while also constraining computational support for language documentation and revitalisation. For communities where these languages encode agricultural knowledge, oral governance, ritual practice, and everyday communication, functional translation tools are not merely a technical convenience but part of equitable access to digital information.

A second, less visible barrier operates upstream of parallel data: multilingual MT systems process text through learned subword vocabularies that allocate capacity unevenly across scripts (Chai et al., 2024; Tonja et al., 2024). When a language’s writing system is poorly represented, the same content can be fragmented into long, unstable token sequences a *tokenization premium* that weakens cross-lingual transfer even when parallel data is available (Blevins et al., 2024). Transliteration can reduce such fragmentation in low-resource Indic settings (Winata et al., 2023), but its effect is not uniform: results are mixed for well-resourced pairs (Tran et al., 2026), and shared-script proximity with Hindi does not prevent copying artifacts or lexical-ambiguity errors (Blaschke et al., 2025). Crucially, a conversion that reduces fragmentation under one tokenizer may increase it under another, making script normalisation a model-dependent intervention rather than a harmless pre-

processing step (Xhelili et al., 2024).

We study this interaction for seven endangered or severely underrepresented Indian languages - Gondi, Mundari, Bhili, Santali, Kui, Garo, and Kokborok, spanning four language families and four scripts. We construct a human-curated, speaker-verified Hindi-English-X tri-parallel corpus of ~ 20 k sentences per language, covering agriculture, health, local governance, education, everyday dialogue, and the cultural data. Native speakers from tribal research institutes and community organisations curated or verified all development and test references. Against this benchmark, we evaluate prompted LLMs (GPT-4o-mini, Llama 3 8B, Mistral 7B), fine-tuned subword MT systems (IndicTrans2, NLLB-600M), and byte-level ByT5 under corpus-native, Devanagari-unified, and Latin-unified script conditions. Figure 1 summarizes the full study design, from the speaker-curated Hindi-English-X benchmark to multi-script MT evaluation and tokenizer-level analysis.

Contributions.

- **A controlled multi-script MT benchmark** for seven endangered Indian languages, with human-verified Hindi-English-X tri-parallel data (~ 20 k sentences each) across six community-relevant domains. To our knowledge, this provides the first MT parallel resource at meaningful scale for Kui, Garo, and Kokborok.
- **Diagnostic metrics and evidence that script effects are tokenizer-mediated.** We propose Script Tokenization Parity (STP) and Script Sensitivity (ScSens) to quantify script-tokenizer interaction. Script preferences reverse between IndicTrans2 and NLLB-600M; byte-level ByT5 reduces ScSens $6.8\times$, implicating vocabulary allocation as the dominant mechanism.
- **A principled community-validity boundary.** Santali/Ol Chiki, despite carrying the highest tokenization cost in the corpus, is excluded from script unification because no lossless conversion is supported, establishing that tokenization parity is necessary but not sufficient.
- **A practical screening workflow:** Compute STP before fine-tuning, validate conversion

with community speakers, and treat script choice as a conditional, model-dependent decision.

2 Related Work

Multilingual Transfer for Low-Resource MT.

Massively multilingual MT shares supervision across languages through common parameters and vocabularies (Johnson et al., 2017; Aharoni et al., 2019), but transfer depends on data balance, typological proximity, and whether high-resource languages dominate the shared space (He and Glass, 2020; Fan et al., 2021; Conneau et al., 2020). These conditions are rarely favourable for endangered languages absent from pretraining, and recent evaluations confirm that current systems fail at this tier for Indian tribal languages (Khanuja et al., 2023; Singh et al., 2025a).

Script Diversity and Tokenization Challenges.

Multilingual tokenizers encode scripts unevenly, allocating dense coverage to high-resource orthographies while fragmenting others into long, unstable sequences (Petrov et al., 2023; Rust et al., 2021). Encoding-length ratios as high as fifteen-fold have been reported across languages (Petrov et al., 2023), with downstream consequences for translation quality; complementary information-theoretic analysis relates tokenisation efficiency to downstream performance (Zouhar et al., 2023).

Script conversion can help when it moves a language closer to a well-covered tokenizer (Zhuang et al., 2025), but gains are mixed (Jaavid et al., 2024), and shared Devanagari with Hindi does not prevent drift or copying errors (Zhang et al., 2023). Our experiments move from tokenizer-level diagnosis to controlled MT experiments, showing that the same script conversion can improve one model while degrading performance in another.

Byte-Level Modeling and Community Evaluation.

Byte-level models remove the learned subword vocabulary entirely. ByT5 processes raw UTF-8 bytes (Xue et al., 2022); a sharp decrease in script sensitivity under ByT5 on identical data implicates vocabulary allocation, not script identity, as the dominant source of variance. We use this as a mechanistic test. Endangered-language MT (Nuha and Jatowt, 2025) also requires evaluation beyond automatic metrics: standard scores can miss semantic drift, cultural-term loss, and safety-relevant ambiguity (Nehrdich et al., 2024).

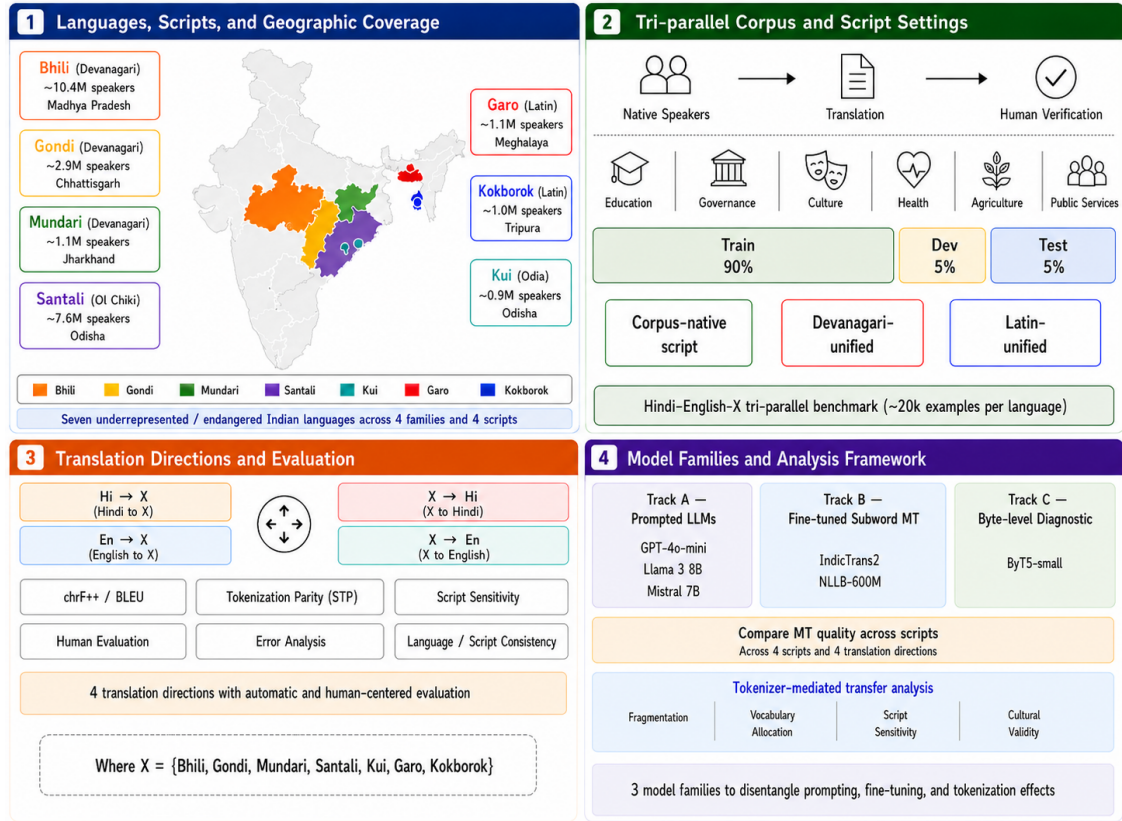


Figure 1: Overview of the benchmark and experimental design. The study covers seven underrepresented Indian languages across four families and four scripts, builds a speaker-verified Hindi-English-X tri-parallel corpus, and evaluates four translation directions under multiple script conditions. Prompted LLMs, fine-tuned subword MT systems, and byte-level models are compared to analyze tokenizer-mediated transfer.

We pair automatic evaluation with native-speaker audits and treat conversion as invalid when it damages linguistic or cultural content. Appendix A extends this discussion with information-theoretic perspectives, and concurrent endangered-language benchmarking.

3 A Grassroots Multi-Script Tri-Parallel Corpus

Motivation. The seven languages in this study are either absent from public MT training corpora or represented only through small, fragmented resources. Samanantar (Ramesh et al., 2022) and BPCC (Gala et al., 2023) include limited Santali coverage; BHEPC (Singh et al., 2025a) contributes Bhili-Hindi-English data; Kokborok, Garo, and Kui have no publicly available parallel corpus of meaningful scale. Web-crawled or pivot-translated alternatives are unreliable at this tier: the digital footprint of these languages is sparse, online orthography is inconsistent, and community-specific vocabulary for kinship, agri-

culture, ritual practice, local governance, and ecological knowledge is either absent from web text or flattened into Hindi and English paraphrases (Bird, 2020; Blasi et al., 2022).

Language Background & Seed Data Creation.

The benchmark covers seven tribal/endangered and severely underrepresented Indian languages spanning distinct regions, families, and writing systems. Bhili is an Indo-Aryan language of western and central India; Gondi and Kui are Dravidian languages associated with central India and Odisha; Mundari and Santali are Austroasiatic languages of eastern India; and Garo and Kokborok are Tibeto-Burman languages of the north-east, including Meghalaya, Assam, and Tripura. Despite active speaker communities, all seven remain weakly represented in public MT resources and multilingual pretraining data. Table 1 summarises the validated script conditions. Gondi, Mundari, and Bhili are represented in our corpus in Devanagari, while Garo and Kokborok are represented in Latin. Kui, represented

Table 1: Languages and script conditions in the Hindi-English-X benchmark. N denotes the corpus-native script condition, ✓ denotes an evaluated script conversion, and “-” denotes a conversion excluded by the linguistic-validity gate. N in a unified-script column indicates that the unified script *is* the corpus-native script, so no conversion is applied. ★ marks languages for which this corpus provides, to the best of our knowledge, the first MT parallel data at meaningful scale.

Lang.	Family	Script	Native	Deva.	Latin
Bhili	Indo-Aryan	Devanagari	N	N	✓
Gondi	Dravidian	Devanagari	N	N	✓
Mundari	Austroasiatic	Devanagari	N	N	✓
Santali	Austroasiatic	Ol Chiki	N	-	-
Kui★	Dravidian	Odia	N	✓	✓
Garo★	Tibeto-Burman	Latin	N	✓	N
Kokborok★	Tibeto-Burman	Latin	N	✓	N

in Odia, provides a non-Devanagari, non-Latin conversion case, while Santali in Ol Chiki is treated separately because validated lossless conversion to Devanagari or Latin was not supported in our pipeline. The corpus contains Hindi-English-X triples across six community-relevant domains: agriculture, health, local governance, education, everyday dialogue, and cultural knowledge. Seed sentences were drawn from public-domain or permissively licensed sources; when one high-resource side was missing, it was produced with IndicTrans2 and verified by a bilingual reviewer, but endangered-language gold references were never model-generated. Additional metadata, morphological complexity profiles (Figure 4, Appendix B), and transliteration validation details appear in Appendices B & C.

Translation Workflow and Quality Assurance.

Approximately five native speakers per language contributed over three months, selected for active language competence, literacy in the corpus script, and domain familiarity. Where possible, we recruited teachers, community literacy workers, and contributors involved in language documentation; compensation followed government daily-wage norms through tribal research institutes. Every translation underwent independent review by a second fluent speaker checking semantic fidelity, script consistency, named entities, numerals, and domain-specific terminology. Flagged sentences were corrected and rechecked; unresolvable cases were excluded. After verification, we applied Unicode NFC normalisation, whitespace and punctuation cleanup, and script-consistency checks. Sentence-level filters removed examples

with fewer than 6 or more than 80 words; near-duplicates were collapsed before splitting. Train, development, and test splits were constructed with strict sentence-level disjointness at an approximately 90/5/5 ratio; overall attrition from seed to final benchmark is 2.3% (Table 6), Appendix B. More details in Appendix E, C and D.

Script Conversion Pipeline. Each language is evaluated in up to three representations: corpus-native, Devanagari-unified, and Latin-unified. Conversions use IndicXlit (Madhani et al., 2023), the Transformer-based neural transliterator released with the Aksharantar dataset, applied with fixed decoding settings. Because a learned transliterator can introduce systematic errors, each condition passes two validity gates that mitigate rather than eliminate conversion error: (i) automatic checks for unmapped characters, round-trip character error, mixed-script artifacts, and proper-noun distortion; and (ii) native-speaker review on a random sample. The corpus-native script condition provides the baseline for each language: Devanagari for Gondi, Mundari, and Bhili, Odia for Kui, Ol Chiki for Santali, and Latin for Garo and Kokborok. Script-conversion effects are always reported relative to this corpus-native baseline.

Santali as a Boundary Case. Santali is evaluated only in corpus-native Ol Chiki. The script encodes Santali-specific phonological distinctions central to community education and identity. Neither Devanagari nor Latin conversion passed our validity gate: no existing library supports Ol Chiki conversion without systematic information loss. Santali carries the highest tokenization premium ($2.49\times$ relative to English on NLLB; Table 10, Appendix G), making it the language for which unification would offer the largest efficiency gain. That it must nonetheless be excluded demonstrates that tokenisation efficiency cannot override community linguistic authority (Appendix C).

4 Experimental Setup

We evaluate four directions: Hindi→X, X→Hindi, English→X, and X→English for every language where X is one of seven tribal languages: Bhili, Gondi, Mundari, Santali, Kui, Garo, or Kokborok.

Model Selection. Our comparison isolates three factors: general multilingual knowledge, task-specific adaptation, and tokenizer-mediated script

Table 2: Adaptation gains from GPT-4o-mini five-shot prompting to the best corpus-native fine-tuned MT system, reported as macro-average chrF++ over seven languages.

Direction	Prompted	Fine-Tuned	Δ
Hi→X	23.4	42.9	+19.5
X→Hi	29.4	53.8	+24.4
En→X	19.9	36.5	+16.6
X→En	30.1	48.8	+18.7

Fine-tuned scores select the best corpus-native system per language–direction from IndicTrans2, NLLB-600M, and ByT5-small using development data. Bhili→Hindi IndicTrans2 is marked with [†] because lexical overlap may inflate automatic metrics.

effects. **Prompted baselines.** GPT-4o-mini (OpenAI et al., 2024), Llama 3 8B-Instruct (Grattafiori et al., 2024), and Mistral 7B-Instruct (Jiang et al., 2023) are evaluated in zero-shot and five-shot settings to measure how far instruction-tuned LLMs reach without target-language supervision. **Subword MT systems.** We fine-tune two models whose tokenizers differ in script bias. IndicTrans2 (Gala et al., 2023) uses a SentencePiece vocabulary shaped by scheduled Indian languages, giving dense Devanagari coverage and NLLB-600M (Costa-Jussà et al., 2022) distributes capacity more broadly, with stronger Latin-script and English representation. If script effects are tokenizer-mediated, the same conversion should produce different and potentially opposite outcomes under these two vocabularies. **Byte-level diagnostic.** ByT5-small (Xue et al., 2022) encodes raw UTF-8 bytes, eliminating learned vocabulary allocation as a source of fragmentation. It serves as a mechanistic control: if script sensitivity drops sharply under ByT5 on identical training data, the subword vocabulary not script identity is the dominant source of variance. All fine-tuned models share the same train/dev/test splits per language-script condition; hyperparameters and compute details appear in Appendix H.

Diagnostic metrics. We introduce three statistics to quantify script–tokenizer interaction; to our knowledge, no prior work formalises these measures for controlled MT evaluation. **Script Tokenization Parity** (STP) measures the encoding cost of representing language L in script condition S relative to its corpus-native script S_{nat} :

$$\text{STP}(L, S, T) = \frac{\text{SFrag}(L, S, T)}{\text{SFrag}(L, S_{\text{nat}}, T)}. \quad (1)$$

Table 3: Script-conversion effects in X→Hindi translation. Scores are Δ_{MT} in chrF++ relative to the corpus-native baseline. Positive values indicate improvement over the native-script condition. Sign differences across IndicTrans2 and NLLB-600M show tokenizer-mediated reversals.

Language	Model	$\Delta_{\text{MT}}^{\text{Deva}}$	$\Delta_{\text{MT}}^{\text{Latin}}$
Gondi	IT2	+0.0‡	+4.3
Gondi	NLLB	+4.1‡	+13.2
Mundari	IT2	−0.9‡	+1.6
Mundari	NLLB	−10.7‡	+0.3
Bhili	IT2	+0.0‡	−23.8
Bhili	NLLB	−4.2‡	−6.5
Kui	IT2	−1.8	−28.0
Kui	NLLB	+19.3	+12.8
Garó	IT2	+6.7	−6.9‡
Garó	NLLB	+12.0	+4.8‡
Kokborok	IT2	−13.0	+0.0‡
Kokborok	NLLB	−3.8	+0.0‡

‡ Unified script identical to the corpus-native script; trained as a separate run, so these values index run-to-run variation rather than a script effect.

Here, $\text{SFrag}(L, S, T)$ is the mean subword token count per sentence for language L under script condition S and tokenizer T , and S_{nat} denotes the corpus-native script. Values below 1 indicate reduced fragmentation relative to the corpus-native script, while values above 1 indicate increased fragmentation. STP re-anchors established subword-fertility measures (Petrov et al., 2023; Rust et al., 2021) to each language’s corpus-native script, so that a value isolates the encoding cost of a script *conversion* rather than absolute fertility. Because reduced fragmentation improves translation only when the resulting subwords are well supported, STP functions as a tokenizer-conditioned screening diagnostic rather than a model-independent quality predictor (Figure 5).

Full STP values appear in Table 10 (Appendix G). **Fragmentation change & MT change,**

$$\begin{aligned} \Delta_{\text{frag}} &= \text{Fert}_{L, S_{\text{nat}}, T} - \text{Fert}_{L, S, T}, \\ \Delta_{\text{MT}} &= \text{chrF}^{++}_{L, S, M} - \text{chrF}^{++}_{L, S_{\text{nat}}, M}. \end{aligned} \quad (2)$$

test whether reduced tokenization cost corresponds to translation gain (Figure 5 in Appendix G). **Script Sensitivity** measures how much a model’s quality varies across script representations of the same content:

$$\text{ScSens}_{M, L} = \max_S \text{chrF}^{++}(L, S, M) - \min_S \text{chrF}^{++}(L, S, M). \quad (3)$$

Lower ScSens indicates greater robustness to script choice.

For $X \rightarrow \text{Hindi}$, ScSens varies only the script representation of the source, while all outputs are evaluated against the same Devanagari Hindi references. The resulting differences therefore reflect variation in the input representation rather than changes in the reference script. For $\text{Hindi/English} \rightarrow X$, references are converted to the corresponding target script and outputs are evaluated directly in that script, without back-conversion. Since chrF++ and BLEU may themselves vary across script representations, comparisons in these target- X directions should be interpreted accordingly.

Evaluation. We report chrF++ (Popović, 2017) as the primary metric and BLEU (Post, 2018) for comparability. Pairwise comparisons use paired bootstrap resampling (Koehn, 2004) ($N=1,000$, $p<0.05$); full 95% confidence intervals appear in Appendix K. Automatic scores are supplemented with speaker-based evaluation (Section 6).

5 Results

Three results emerge from the controlled comparison. Prompting fails for this language tier; script unification is not a stable intervention but a tokeniser-dependent one; and removing the learned vocabulary collapses most cross-script variance, locating the mechanism in subword allocation rather than script identity.

Prompting Is Not Enough. Prompted LLMs produce unstable output across the seven languages, establishing a baseline against which the value of community-curated parallel data can be measured. The full 0-shot and 5-shot prompting results for all Native, Devanagari, and Latin script settings are reported in Tables 14-25. GPT-4o-mini achieves its highest five-shot chrF++ of 56.0 on Bhili $\rightarrow$ Hindi, where extensive lexical and script overlap with Hindi inflates surface similarity, but averages only 29.4 across $X \rightarrow \text{Hindi}$ (Table 15). Llama 3 and Mistral frequently substitute Hindi or English for the requested target, a language-routing failure observed for Bhili in prior work (Singh et al., 2025a) and confirmed here across all seven languages. Full prompting results appear in Appendix I.

Fine-Tuning Narrows the Gap. Table 2 summarises the adaptation gain. Fine-tuned systems

Table 4: Script sensitivity in $X \rightarrow \text{Hindi}$ translation. Lower ScSens indicates greater robustness to script choice; ByT5 reduces average sensitivity from 14.2 to 2.1 chrF++. Santali is excluded because lossless Ol Chiki conversion was not supported.

Language	IT2	NLLB	ByT5	IT2/ByT5
Gondi	4.3	13.2	2.8	1.5
Mundari	2.5	11.0	2.0	1.2
Bhili	23.8	6.5	2.4	9.9
Kui	28.0	19.3	3.1	9.0
Garro	13.6	12.0	1.6	8.5
Kokborok	13.0	3.8	0.7	18.6
Average	14.2	11.0	2.1	6.8

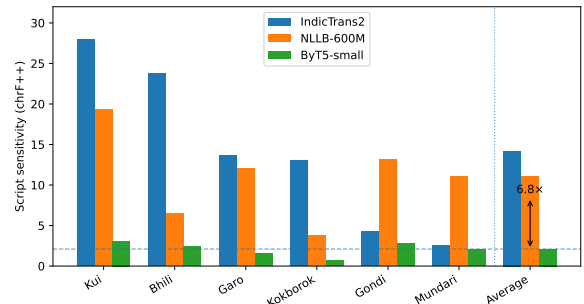


Figure 2: Script sensitivity in $X \rightarrow \text{Hindi}$ translation. Lower ScSens indicates greater robustness to script choice; ByT5 reduces average sensitivity by 6.8 $\times$, indicating that most cross-script variance comes from learned subword vocabulary allocation.

improve over prompted GPT-4o-mini five-shot by 16–25 chrF++ across all four directions, with the largest gain in $X \rightarrow \text{Hindi}$ (+24.4), where encoding the endangered-language source is the bottleneck. The finetuning results for Native, Devanagari, and Latin script settings are reported in Tables 5, 27 & 28. The reverse directions are harder, they require generation in languages with minimal pretraining support but still gain 17-20 points. Roughly 20k in-domain triples suffice to move these languages from unreliable prompting to functional translation. Table 5 reports full corpus-native fine-tuning grid across languages, directions, and model families. Unified-script results are in Appendix J; we flag Bhili $\rightarrow$ Hindi there because speaker audit finds lexical overlap and copying artifacts that inflate automatic scores (Appendix F) and training details are in Appendix H & J.

Script Effects Are Tokenizer-Dependent. The same script conversion can have opposite effects across models, so script alone cannot explain the observed differences. We analyse script effects

within each model while holding the architecture, base checkpoint, data splits, training configuration, and decoding fixed and varying only the script representation of the same content (Section 4; Appendix H). Thus, each Δ_{MT} measures a fixed model’s response to script representation. Since the model families also differ in architecture and pretraining, we use *tokenizer-mediated* to identify learned vocabulary allocation as a supported contributor rather than the exclusive cause.

Table 3 reports Δ_{MT} for $X \rightarrow \text{Hindi}$ from the fine-tuned results in Appendix J, with positive values indicating improvement over the corpus-native baseline. Latin conversion degrades Kui by 28.0 chrF++ under IndicTrans2 but improves it by 12.8 under NLLB-600M; Devanagari is near-neutral under IndicTrans2 (−1.8) but strongly beneficial under NLLB-600M (+19.3). Devanagari also improves Garo under both systems but harms Kokborok under IndicTrans2 (−13.0). These reversals are not explained by lexical copying: Kui has only 0.001 word-level overlap with Hindi and substantially lower subword overlap than Bhili across all three tokenisers (Table 11), yet its conversion still flips sign across models. Bhili $\rightarrow$ Hindi, where high overlap ($ov=0.608$ under IndicTrans2) may inflate automatic scores, is treated separately as a confound and marked with [†] (Appendix F). Under these controlled comparisons, the reversals therefore point to model-specific vocabulary allocation rather than script identity alone (Figure 3).

Tokenization statistics further support this account (Table 10; Appendix G). For Kui under IndicTrans2, Devanagari reduces fertility from 6.05 to 2.39 (STP=0.396), mean sentence length from 71.9 to 28.5 pieces, and raises observed vocabulary utilisation from 0.26% to 9.33% (Table 13), yet yields only $\Delta_{MT}=-1.8$, compared with +19.3 under NLLB-600M. Thus, reduced fragmentation alone is insufficient: vocabulary overlap and subword quality modulate its translation effect. Figure 5 similarly shows that reduced fertility correlates with chrF++ gains under NLLB-600M but not IndicTrans2.

Byte-Level Models Localize the Mechanism.

If script effects primarily arose from properties of the scripts themselves, they should persist without a learned subword vocabulary. Instead, mean $X \rightarrow \text{Hindi}$ ScSens falls from 14.2 chrF++ for IndicTrans2 and 11.0 for NLLB-600M to 2.1 for ByT5-

small, a $6.8\times$ reduction relative to IndicTrans2 (Table 4; Table 29; Figure 2). The largest reductions occur for Kokborok, Bhili, and Kui, where subword-model sensitivity is highest. Residual ByT5 sensitivity can reflect byte-length asymmetry, but UTF-8 byte units are deterministic, lossless, and shared across scripts; thus, high ByT5 fertility represents a sequence-length cost rather than the rare, script-specific fragmentation associated with vocabulary-exhaustion failures in subword models (Xue et al., 2022). This contrast supports learned vocabulary allocation, rather than sequence length alone, as a major contributor to cross-script variance.

Because chrF++ is a character n -gram metric, some measured ScSens could in principle be metric-induced. The $X \rightarrow \text{Hindi}$ comparison limits this confound: only the source representation varies, while every condition is scored against the same unchanged Devanagari Hindi references. Under this controlled setup, ByT5 exhibits only 2.1 chrF++ mean ScSens, providing an empirical estimate of residual sensitivity; the roughly 12 chrF++ gap between IndicTrans2 and ByT5 therefore cannot plausibly be explained by the metric alone.

The pattern persists at larger capacity: for Kui, the highest-ScSens case under ByT5-small, fine-tuned ByT5-base ($\sim 580M$) varies by only 0.45 chrF++ across source scripts for $X \rightarrow \text{Hindi}$ (Appendix J; Table 26). Nor is the reduction explained by score range: after normalising ScSens by each model’s mean chrF++, ByT5 remains roughly $6.5\times$ less sensitive than IndicTrans2, while Kui—its lowest-scoring $X \rightarrow \text{Hindi}$ case—still has the largest ScSens among the ByT5 languages (Table 30). Hindi $\rightarrow$ Kui nevertheless favours Latin under ByT5-base, showing reduced source-side script sensitivity without guaranteeing target-side invariance. Restricting ScSens to the seven validated conversions, excluding cases where the unified and corpus-native scripts coincide, likewise preserves the pattern: mean sensitivities are 12.9, 9.2, and 2.0 chrF++ for IndicTrans2, NLLB-600M, and ByT5-small, respectively. Together, the capacity, score-range, reference, and script-grid controls support learned vocabulary allocation as a major contributor to the observed script sensitivity.

Fragmentation Alone Does Not Explain Quality. Santali and Kui provide the strongest coun-

terexample to a naïve “shorter-is-better” account. Santali has the highest tokenization premium in the corpus ($2.49\times$ vs. English on NLLB; Table 10), yet achieves the highest corpus-native chrF++ among the seven languages in both $X\rightarrow\text{Hindi}$ (82.2, IT2) and $\text{Hi}\rightarrow X$ (65.8, NLLB). Kui has lower native fertility but is the most fragile generation case: $\text{En}\rightarrow\text{Kui}$ peaks at only 17.7 (ByT5), with qualitative analysis revealing vocabulary exhaustion and critical failures. This contrast separates fragmentation from effective subword coverage: Santali benefits from consistent Ol Chiki representation in IndicTrans2’s pretraining, whereas Kui suffers from poor Odia-script subword support despite lower raw token counts.

The pattern remains model-dependent across all four translation directions. Pooling the 28 valid script-conversion cases per model, reduced fragmentation is moderately associated with translation gain under NLLB-600M ($\rho = +0.56$; slope $+5.9$, 95% CI $[-2.2, +16.7]$), but only weakly under IndicTrans2 ($\rho = +0.14$; slope -0.4 , 95% CI $[-4.1, +15.4]$), where slopes denote chrF++ change per subword token per word; both intervals include zero. These pooled correlations differ from the $X\rightarrow\text{Hindi}$ -only values in Figure 5 ($\rho = +0.61$ for NLLB-600M and -0.21 for IndicTrans2) because they include all four directions. Overall, lower fragmentation can help but does not consistently imply better MT quality across models (Appendix G; Table 12).

6 Qualitative Error Analysis and Human Evaluation

Automatic metrics capture aggregate trends but cannot determine whether an output is acceptable for the communities it is meant to serve. A translation may obtain partial chrF++ overlap while drifting into a neighbouring language, flattening a culturally specific term, or producing unreadable script artifacts. We therefore complement automatic evaluation with speaker-accountable error analysis (Nimo et al., 2025) across the full experimental grid.

Annotation Protocol. Expert linguists annotated outputs using an eight-class error taxonomy (E1–E8; Appendix M) and a five-point MQM-inspired severity scale collapsed into critical, major, and minor bands. Two linguists independently annotated each segment, with disagreements adjudicated by a senior speaker. Annotation guidelines,

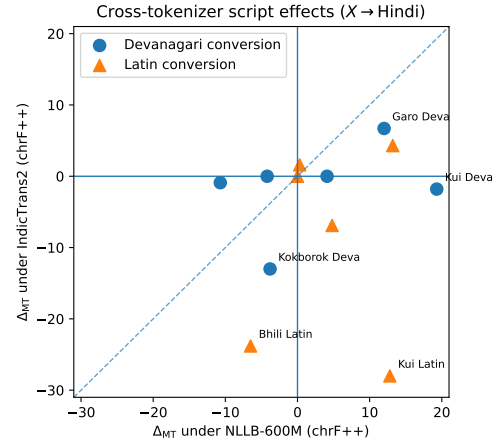


Figure 3: Cross-tokenizer script effects in $X\rightarrow\text{Hindi}$ translation. Each point is a language–script conversion plotted by Δ_{MT} under NLLB-600M and IndicTrans2. Points in opposite-sign regions show tokenizer-mediated reversals, where the same conversion helps one model but hurts the other.

agreement statistics, and full distributions are provided in Appendices L and N. The audit covers a stratified sample of 259 corpus-native IndicTrans2 outputs across languages and translation directions, rather than every model-script condition; comparative samples from the other systems and unified-script conditions are inspected separately (Appendix N). The human analysis therefore complements the aggregate automatic evaluation; automatic-score gains alone are not treated as evidence of acceptable translation.

Error landscape. The error profile varies sharply by language family, script condition, and model. Devanagari-native languages mainly show semantic adequacy errors (E1), but differ in secondary failures: Bhili shows target-language drift toward Gujarati (E5), while Mundari shows morphological flattening (E6) in agglutinative verb clusters. Latin-native Garo and Kokborok suffer under IndicTrans2’s Devanagari-heavy tokenizer but improve under NLLB-600M, whose vocabulary provides broader Latin coverage. Under Devanagari unification, Garo improves while Kokborok degrades under IndicTrans2, matching the Δ_{MT} reversals in Table 3. Kui concentrates the most severe failures: garbled mixed-script output and malformed token sequences (E8), together with routing toward Standard Odia (E5). Santali provides the opposite case: despite the highest tokenisation premium, it remains strong in corpus-native Ol Chiki, showing that effective subword

Table 5: Corpus-native fine-tuning results across directions and languages. Each cell reports BLEU/chrF++ in the native-script condition. † marks Bhili→Hindi, where lexical overlap may inflate automatic scores (Appendix F).

Dir.	Model	Gondi	Mundari	Bhili	Santali	Kui	Garo	Kokborok
Hi→X	IT2	11.04/36.13	10.07/22.64	49.23/54.06	67.05/64.77	2.35/23.04	19.13/33.43	34.39/37.34
	NLLB	11.98/30.38	18.53/43.45	34.93/53.90	46.14/65.76	1.42/14.70	2.39/21.83	12.88/44.24
	ByT5	11.81/36.36	17.36/41.45	17.26/41.97	26.54/46.06	2.07/21.58	5.78/33.39	14.44/43.00
X→Hi	IT2	17.29/23.00	28.04/29.76	84.24/79.15†	82.93/82.19	23.94/41.88	19.13/30.33	19.18/48.63
	NLLB	17.38/33.46	30.19/49.23	40.33/56.69	49.45/64.12	3.21/16.32	11.11/27.24	22.68/43.25
	ByT5	23.71/45.35	25.50/45.87	35.93/54.28	28.32/43.78	5.77/22.27	12.10/30.46	11.44/29.96
En→X	IT2	6.27/30.98	13.65/21.53	12.98/31.95	49.32/44.91	0.90/9.74	19.30/32.60	10.07/30.54
	NLLB	5.61/21.07	11.06/33.52	11.57/26.92	31.35/50.33	1.55/14.73	10.62/39.69	20.94/51.06
	ByT5	5.01/23.11	7.81/26.90	5.52/21.95	13.09/27.62	1.59/17.72	13.02/39.43	22.08/49.95
X→En	IT2	64.35/49.98	33.52/38.78	12.30/47.63	63.00/59.96	20.38/39.80	30.74/28.96	15.41/51.90
	NLLB	5.87/27.26	24.08/45.88	10.70/35.22	25.05/49.09	4.66/22.71	25.14/46.74	26.84/48.45
	ByT5	10.32/31.41	17.04/37.77	13.95/36.17	12.31/31.13	5.98/24.93	22.38/43.62	20.54/41.61

coverage matters more than raw sequence length (examples in Figures 8- 11, Appendix M & O).

Direction- and Model-Specific Failures. Error severity varies by direction and model family. Into-X directions (En→X, Hi→X) concentrate critical errors because decoding occurs in sparsely covered vocabulary regions. En→X is hardest, with garbled Kui output (E8), Bhili language drift (E5), and semantic hallucination in Garo and Kokborok (E1). X→Hindi and X→English are more reliable because decoding occurs in high-resource languages, while Hi→X lies between these extremes, benefiting from Hindi–X lexical proximity for Devanagari-script languages but degrading for distant pairs such as Hindi→Garo and Hindi→Kokborok. Across script conditions, IndicTrans2 is strongest for Devanagari-proximate languages but fails on script-distant pairs, particularly with E8 for Kui and E5 for Bhili and Latin-native languages under Devanagari conversion. NLLB-600M avoids most garbled output but shows more undertranslation (E7) and Odia- or Hindi-influenced paraphrases when target vocabulary is weak. ByT5 nearly eliminates E8, supporting its association with learned subword vocabulary limitations, but shows higher rates of E7 and semantic-adequacy errors (E1). Thus, failure type is model-specific, while its frequency varies with the script–tokenizer interaction.

Cultural & Lexical Fidelity. Across languages, culturally grounded terms are often flattened into generic Hindi or English equivalents: agricultural vocabulary specific to tribal farming, kinship terms encoding local social relations, and governance terms rooted in community institu-

tions. A related failure arises when endangered-language words share surface orthography with high-resource vocabulary but carry different meanings. These cases are annotated as semantic-adequacy errors when the meaning is reversed, and as cultural lexical substitutions when a community-specific term is replaced by a generic high-resource equivalent. Such errors can receive partial automatic credit while still damaging speaker trust and documentation value (Bird, 2020; Mager et al., 2023). Full per-language error distributions, and representative examples appear in Appendices N, O; cultural aspects of the dataset are discussed in Appendix P.

7 Conclusion

We introduced a speaker-curated Hindi–English–X tri-parallel benchmark for seven endangered Indian languages spanning four families and four scripts, providing first-at-scale MT parallel data for Kui, Garo, and Kokborok. Controlled multi-script experiments show that fine-tuning on roughly 20k triples improves chrF++ by 16-25 points over prompting; script preferences reverse across tokenizers; and byte-level ByT5 reduces script sensitivity by 6.8×, locating much of the variance in learned vocabulary allocation. Santali shows that tokenization efficiency alone is insufficient: script conversion must also be technically supported, linguistically valid, and community-acceptable. The practical recommendation is to measure tokenization parity before fine-tuning, validate script representations with speakers, and treat script choice as a conditional modelling decision rather than routine preprocessing.

Limitations

This work studies machine translation as one component of a broader language-technology ecosystem for endangered-language communities. MT can support access to information, corpus development, and language documentation, but language vitality depends on speaker communities, intergenerational transmission, schools, cultural institutions, and locally governed revitalisation efforts. We therefore frame translation technology as assistive infrastructure that should serve, not replace, community-led preservation work. The empirical scope is deliberately controlled: seven Indian languages, three model families, four translation directions, and the script conditions supported and validated in our pipeline. The reported effect sizes should be read as setting-specific. Since the three model families also differ in architecture and pretraining, cross-model comparisons cannot attribute these differences to tokenisation alone; our controlled within-model results instead support learned vocabulary allocation as an important contributor to script sensitivity. The transferable contribution is methodological: measure tokenizer behaviour before fine-tuning, validate script representations with speakers where possible, and inspect culturally and semantically consequential errors that automatic metrics may understate. Future work should extend this framework to document-level MT, speech interfaces, dialectal variation, and terminology-rich deployment in domains such as health, education, agriculture, and local governance. The diagnostic workflow—STP measurement, community-accountable evaluation, and conditional script decisions—is language-agnostic and reusable beyond the seven languages studied here (Appendix Q).

Ethical Considerations

This work concerns languages whose writing practices are tied to education, literature, identity, and community authority. Participants were informed about the project goals, data collection process, and intended research and documentation uses; consent procedures were provided in locally understandable languages, and informed verbal or written consent was obtained. Contributors were recruited through community-facing channels, compensated according to local daily-wage norms, and involved as linguistic experts through community linguists, educators, and local language practition-

ers. The Devanagari- and Latin-unified conditions are used only as experimental representations for measuring tokenizer-mediated transfer, not as recommendations for script choice or orthographic standardisation. We do not ask speakers to adapt their languages to existing models; we ask whether MT systems can serve the representations that speakers and communities consider valid. All development and test references are human-curated, post-edited, or speaker-verified, and no model-generated endangered-language text is used as a gold reference. Santali is evaluated only in Ol Chiki because lossless conversion to Devanagari or Latin was not supported in our pipeline. The corpus is intended as a documentation-oriented research resource for language preservation, education, and community-centered speech technology, and should not be used for forced script normalisation, community misrepresentation, or deployment without speaker review and appropriate local governance.

Acknowledgments

This work was carried out under the AdiVaani initiative of the Ministry of Tribal Affairs, Government of India, supporting the development of language technologies and digital resources for tribal and endangered Indian languages. We gratefully acknowledge the Ministry’s support to IIT Delhi, which enabled the research, language-resource development, community engagement, and computational infrastructure underlying this work. We also thank IIT Delhi HPC for providing the computational resources used in our experiments. We thank Ms. Ranjana Chopra, Shri Anant Prakash Pandey, Ms. Deepali R. Masirkar, and Shri Ganesh Nagarajan (Ministry of Tribal Affairs); Prof. Vivek Kumar and Prof. Parag Singla (IIT Delhi) for their guidance and institutional support.

We gratefully acknowledge the Tribal Research Institutes (TRIs), local institutions, language experts, and community partners who facilitated language-resource development and engagement with the seven language communities represented in this work: Bhili in Madhya Pradesh, Mundari in Jharkhand, Gondi in Chhattisgarh, Santali and Kui in Odisha, Kokborok in Tripura, and Garo in Meghalaya. Their support was instrumental in facilitating community engagement, language-data collection, linguistic validation, and coordination with native speakers and language experts. We

thank all language experts and community contributors for their linguistic knowledge, care, and sustained effort in developing and validating the resources used in this study. We also thank Amit Kumar, Avinash Soy and Mohammad Fazal (IIT Delhi); Mithun Chandravanshi, Praveen, Nicolette, and Anuj Tandon (Ministry of Tribal Affairs) for their support in coordinating the broader project activities. We further acknowledge the researchers, students, developers, and technical teams who contributed to the recording, annotation, validation, data management, and computational infrastructure supporting this work.

Most importantly, we thank the speakers and communities of Bhili, Mundari, Gondi, Santali, Kui, Kokborok, and Garo for contributing their language, time, and linguistic knowledge. Their participation is fundamental to the development of community-grounded resources for endangered and underrepresented Indian language technology.

References

- Roei Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.
- Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.
- Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684.
- Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the worlds languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.
- Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 10822–10837.
- Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. 2024. Tokenization falling short: On subword robustness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1582–1599.
- Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096.
- 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.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *Journal of Machine Learning Research*, 22(107):1–48.
- Jay P Gala, Pranjal A Chitale, Varun Gumma, Sumanth Doddapaneni, Kumar M Aswanth, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Tianxing He and James Glass. 2020. Negative training for neural dialogue response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2044–2058.
- J Jaavid, Raj Dabre, M Aswanth, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models via romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615.

- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). Preprint, arXiv:2310.06825.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vi  gas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. [IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. [MuriL: Multilingual representations for indian languages](#). Preprint, arXiv:2103.10730.
- Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of nlp technology: A case study for indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.
- Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations*, pages 66–71.
- Tomasz Limisiewicz, Ji  r   Balhar, and David Mare  ek. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681.
- Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2023. Aksharantar: Open indic-language transliteration datasets and models for the next billion users. In *Findings of the association for computational linguistics: Emnlp 2023*, pages 40–57.
- Manuel Mager, Elisabeth Maier, Katharina von der Wense, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897.
- Benjamin Muller, Antonios Anastasopoulos, Beno  t Sagot, and Djam   Seddah. 2021. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462.
- Sebastian Nehrlich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: Byt5-sanskrit, a unified model for sanskrit nlp tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751.
- Charles Nimo, Shuheng Liu, Amy Z Chen, Ramaravind Kommiya Mothilal, and Michael L Best. 2025. Community-driven data practices for advancing ethical and equitable ai in low-resource language contexts. In *Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 844–847.
- Ulin Nuha and Adam Jatowt. 2025. Exploring nlp benchmarks in an extremely low-resource setting. *arXiv preprint arXiv:2509.03962*.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). Preprint, arXiv:2410.21276.
- Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.
- Maja Popovi  . 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.
- Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on*

- machine translation: Research papers*, pages 186–191.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725.
- Pooja Singh, Shashwat Bhardwaj, Vaibhav Sharma, and Sandeep Kumar. 2025a. Leveraging the cross-domain & cross-linguistic corpus for low resource nmt: A case study on bhili-hindi-english parallel corpus. *arXiv preprint arXiv:2511.00486*.
- Pooja Singh, Sandeep Chatterjee, Gullal S Cheema, Amrit Singh Bedi, Tanmoy Chakraborty, Sandeep Kumar, and Ankita Shukla. 2025b. Findings of the mmloso 2025 shared task on machine translation into tribal languages. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pages 121–129.
- Atnafu Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. Nlp progress in indigenous latin american languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6972–6987.
- Van-Hien Tran, Huy Hien Vu, Hideki Tanaka, and Masao Utiyama. 2026. Representation-aware prompting for zero-shot marathi text classification: Ipa, romanization, repetition. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 436–443.
- Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.
- Orgest Xhelili, Yihong Liu, and Hinrich Schütze. 2024. Breaking the script barrier in multilingual pre-trained language models with transliteration-based post-training alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11283–11296.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.
- Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Dont trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927.
- Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2025. Enhancing cross-lingual transfer through reversible transliteration: A huffman-based approach for low-resource languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16299–16313.
- Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207.

Appendix Overview

A Supplementary Related Work

B Language Metadata and Statistics

C Corpus Construction and Transliteration

D Translation Guidelines and Verification

E Annotation and Data Collection

F Bhili Lexical Overlap Analysis

G Tokenization Metrics and Vocabulary Overlap

H MT Systems, Training, and Compute

I Prompt Templates and Prompting Results

J Fine-Tuned Results Across Scripts

K Tokenizer-Mediated Effects

L Error Annotation and Verification

M Human Evaluation Error Taxonomy

N Error Patterns Across Settings

O Representative Translation Errors

P Cultural Representation and Governance

Q Generalization Beyond the Benchmark

R Dataset Scale and Robustness

A Supplementary Related Work

This appendix extends Section 2 with four lines of work that contextualise our contributions within the broader landscape of tokeniser-aware multilingual NLP.

Information-theoretic perspectives on tokenizer disparity. Kudo (2018) and Zouhar et al. (2023) frame subword segmentation as a noisy-channel compression problem: a tokeniser trained on a skewed language distribution allocates more bits per character to under-represented scripts as a direct consequence of its training objective. Limisiewicz et al. (2023) extend this to multilingual inference, showing that scripts absent from a model’s high-frequency vocabulary region incur comprehension penalties independent of the downstream task. These results provide the theoretical basis for the fragmentation-mediated transfer hypothesis tested in Section 5: script conversion can improve translation quality precisely when it moves a language into a lower-entropy region of the tokeniser’s compression space, and the direction of that improvement is contingent on the tokeniser, not the script.

Architectural responses to the vocabulary bottleneck. Rather than normalising scripts post hoc (Sennrich et al., 2016; Kudo and Richardson, 2018), Blevins et al. (2024) train cross-lingual expert models that partition multilingual capacity by language group, and Chang et al. (2024) show across 250 languages that multilingual data helps low-resource languages only up to a data-balance threshold, after which vocabulary contention drives quality below monolingual baselines. These findings confirm that the vocabulary bottleneck has consequences well beyond the script-conversion setting studied here; our ByT5

diagnostic provides a complementary lens by removing the learned vocabulary entirely rather than redesigning it.

Transliteration in multilingual pretrained models. Muller et al. (2021) show that romanisation via UROMAN enables zero-shot transfer to unseen scripts by collapsing character sequences into the Latin subword region. Xhelili et al. (2024) extend this to post-training alignment, and Zhuang et al. (2025) introduce Huffman-based reversible transliteration that maximises subword overlap while preserving round-trip fidelity. These results confirm that script normalisation is an effective lever for improving tokeniser coverage but that its benefit is mediated by the target model’s vocabulary, the dependency that motivates our model-conditioned experimental design.

Concurrent endangered-language benchmarking. The MMLoSo 2025 shared task (Singh et al., 2025b) evaluates MT into four tribal languages with 20k training pairs per language. Our corpus was collected independently through the pipeline described in Section 3, is not derived from that release, and additionally covers Kokborok, Garo, and Kui. Our work also differs in experimental focus: rather than benchmarking systems in a fixed setting, we use a controlled multi-script design and a byte-level diagnostic to examine tokenizer-mediated cross-script variation. Khanuja et al. (2023) quantify ecosystem-level support gaps for tribal Indian languages; Tonja et al. (2024) document parallel gaps for indigenous Latin American languages, confirming that the structural disparities motivating this work extend well beyond the Indian subcontinent.

B Language Metadata, Statistics & Morphological Profiles

Table 7 consolidates the metadata, script conditions, and corpus statistics needed to interpret both the tokenisation analysis (Section 5) and the qualitative error patterns (Section 6).

The seven languages span four families and four corpus scripts, while also differing substantially in morphological profile (Table 8; Figure 4). Garo and Kui have the highest lexical diversity (TTR 0.298 and 0.274, respectively; MATTR above 0.83), consistent with their difficulty in target-language generation: a large type inventory combined with sparse subword coverage increases the

Table 6: Corpus construction pipeline and final split sizes. The table tracks sentence counts from the initial Hindi-English seed pool through target-language translation, speaker verification, filtering, and released train/dev/test splits. Filtering applies length constraints, script-consistency checks, and near-duplicate removal; attrition reports the percentage reduction from seed to final examples.

Language	Seed	Translated	Verified	Final	Final splits			
					Train	Dev	Test	Attrit.
Bhili	20,120	20,020	19,820	19,680	17,712	984	984	2.2%
Gondi	20,080	19,990	19,850	19,640	17,676	982	982	2.2%
Mundari	20,210	20,090	19,890	19,750	17,776	987	987	2.3%
Santali	20,350	20,230	19,960	19,870	17,882	994	994	2.4%
Kui	20,420	20,290	20,050	19,940	17,946	997	997	2.4%
Garó	20,170	20,070	19,860	19,720	17,748	986	986	2.2%
Kokborok	20,260	20,140	19,920	19,800	17,820	990	990	2.3%
Total	141,610	140,830	139,350	138,400	124,560	6,920	6,920	2.3%

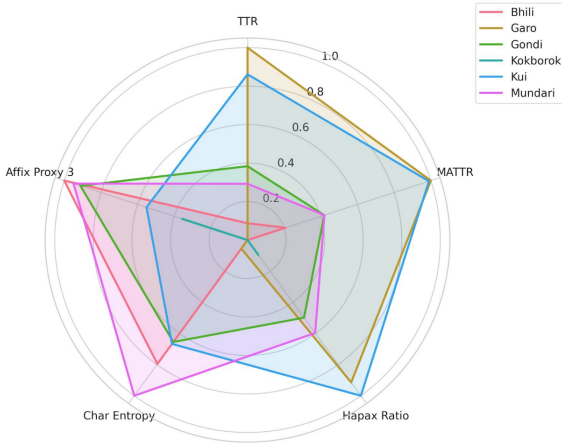


Figure 4: Normalized morphological-complexity profiles for the target-language corpora. Each axis is scaled to $[0, 1]$ to compare lexical diversity, moving-average diversity, hapax rate, character entropy, and affix-proxy features across languages. The profile highlights that Garo and Kui have the highest lexical diversity, while Bhili and Mundari show stronger affixal complexity.

risk of degenerate or hallucinated output, matching the E8 and E1 patterns observed in Section 6. Bhili and Mundari show the highest affix-proxy values (0.248 and 0.241 at $k=3$), reflecting the morphological complexity that appears as E6 flattening when agglutinative verb clusters are segmented into opaque Devanagari units. Speaker counts in Table 7 are approximate lower-bound estimates from the 2011 Census of India. Because census reporting aggregates some tribal and regional language names, these figures should be treated as floor estimates. ISO 639-3 codes and genealogical classifications follow Glottolog.²

²Speaker counts are approximate 2011 Census lower-

Santali is the only language in this set included in IndicTrans2’s pretraining distribution as one of the 22 scheduled Indian languages; the other six are zero-shot for that model before fine-tuning. None of the seven appears in NLLB-600M’s official supported-language list, although its 200-language vocabulary provides partial Devanagari-script coverage. ByT5 uses no language-specific vocabulary and processes all seven through raw UTF-8 bytes. The split ratio is approximately 90/5/5 per language, yielding roughly 18k training, approx 1k development, and approx 1k test sentences; exact per-language counts are reported in Table 6.

Corpus construction attrition. Table 6 traces sentence counts through each stage of the construction pipeline. The initial seed pools contain roughly 20–20.5k sentences per language and yield about 19.6–19.9k filtered examples after translation, speaker verification, cleaning, and deduplication. Overall attrition from seed collection to the final released benchmark is 2.3%, with per-language rates ranging from 2.2% to 2.4%. The small differences between seed, verified, and final counts indicate that quality control primarily refined the corpus rather than substantially reducing it.

bound estimates; ISO 639-3 codes and language-family classifications follow Glottolog: Gondi <https://glottolog.org/resource/languoid/id/gond1265>; Mundari <https://glottolog.org/resource/languoid/id/mund1320>; Bhili <https://glottolog.org/resource/languoid/id/bhil1251>; Santali <https://glottolog.org/resource/languoid/id/sant1410>; Kui <https://glottolog.org/resource/languoid/id/kui1252>; Garó <https://glottolog.org/resource/languoid/id/garo1247>; Kokborok <https://glottolog.org/resource/languoid/id/kokb1239>.

Table 7: Full language metadata and target-side corpus statistics for the Hindi–English–X benchmark. Speaker counts are approximate lower-bound estimates from the 2011 Census of India. ISO 639-3 codes and language-family classifications follow Glottolog.

Lang.	ISO	Family	Speakers	Region	Avg. w.	Avg. ch.	Train tok.	Test tok.	Train type	Test type
Bhili	bhb	Indo-Aryan	10.4M	Western India	17.39	67.93	330,774	17,123	45,482	6,123
Garó	grt	Tibeto-Burman	1.15M	Meghalaya–Assam	11.24	77.29	213,773	11,032	53,205	5,746
Gondi	gon	Dravidian	2.98M	Central India	9.47	42.87	179,942	9,525	33,465	4,193
Kokborok	trp	Tibeto-Burman	0.92M	Tripura–Assam	17.11	99.03	325,370	17,152	31,058	4,503
Kui	kxu	Dravidian	0.94M	Odisha interior	11.88	60.89	225,889	11,746	59,147	6,288
Mundari	unr	Austroasiatic	1.13M	Eastern India	15.22	69.37	289,034	15,312	45,691	5,655
Santali	sat	Austroasiatic	7.37M	Eastern India	20.97	113.05	398,019	21,324	18,058	2,887

Table 8: Morphological complexity of the target-language corpora. The table reports lexical diversity, moving-average diversity, hapax rate, character entropy, and suffix-based affix proxies; high-diversity languages such as Garó and Kui correspond to harder target-language generation settings.

Lang.	TTR	MATTR	Hapax	Char H	Afx ₃	Afx ₄
Bhili	.139	.703	.633	5.09	.248	.523
Garó	.298	.835	.723	4.49	.094	.223
Gondi	.191	.738	.682	4.97	.235	.483
Kokborok	.124	.668	.642	4.44	.149	.345
Kui	.274	.834	.731	4.98	.179	.347
Mundari	.175	.738	.692	5.25	.241	.494

B.1 Dataset Availability

We introduce one of the first large-scale tri-parallel resources for seven extremely low-resource Indian tribal languages, developed through community-driven translation and verification. The Hindi–English–X corpus, evaluation splits, transliteration resources, fine-tuning and evaluation code, and diagnostic scripts (STP, ScSens, Δ Frag) will be released under a research license for academic and non-commercial use.

C Corpus Construction and Transliteration Validation

The tri-parallel corpus was constructed primarily through human translation by native speakers. Model assistance was used only for the high-resource Hindi–English side when a seed sentence was available in one language, and for producing transliterated representations used in unified-script experiments. These automatic steps are recorded in sentence-level provenance metadata; endangered-language gold references were not model-generated.

Source seeds and high-resource sides. Seed sentences were drawn from public-domain or per-

missively licensed materials in Hindi or English across the six study domains. Where a seed was available in only one language, IndicTrans2 produced a draft of the missing high-resource side; a bilingual reviewer then verified the draft against the original before the pair entered the pipeline. This workflow applies exclusively to the Hindi–English pair. Endangered-language references are always human-curated.

Endangered-language translation. Each verified Hindi–English pair was assigned to a native-speaker translator for the target language, who produced the translation in the designated corpus script (Devanagari for Gondi, Mundari, and Bhili; Odia for Kui; Ol Chiki for Santali; Latin for Garó and Kokborok). Translators were selected for active language competence, literacy in the corpus script, and familiarity with the target domains. Translation guidelines, piloted with senior community members before full-scale collection, are documented in Appendix D.

Script transliteration. For the Devanagari-unified and Latin-unified experimental conditions, script conversion was performed using IndicXlit (Madhani et al., 2023), the open-source Transformer-based transliteration model released with the Aksharantar dataset, which supports transliteration across major Indic scripts including Devanagari, Odia, Bengali, Gurmukhi, and Latin (ISO 15919). IndicXlit was selected because it is openly available, reproducible without proprietary tooling, and supports round-trip conversion for automatic quality measurement. Because IndicXlit is a learned model rather than a fixed rule table, its mappings are not inherently auditable; we therefore rely on the two-gate validation protocol below to detect conversion errors. All outputs were post-processed for Unicode NFC normalisation and whitespace consistency.

Transliteration validation. For each non-trivial conversion pair, we applied a two-stage validation protocol: Kui from Odia to Devanagari and Latin; Gondi, Mundari, and Bhili from Devanagari to Latin; and Garo and Kokborok from Latin to Devanagari. First, four automatic quality metrics were computed over the full converted corpus: unmapped-character rate, round-trip character error rate, mixed-script rate, and proper-noun distortion rate. Second, manually inspected stratified samples across all six domains, focusing on character coverage, reversibility, readability, proper names, and obvious meaning-changing distortions. A conversion was retained only when the automatic checks passed the predefined thresholds reported in the dataset documentation, and manual inspection found no systematic conversion artifacts.

Santali Ol Chiki was the sole case where conversion failed this gate. No existing transliteration library or model including IndicXlit (Madhani et al., 2023) supports Ol Chiki→Devanagari or Ol Chiki→Latin, because the script encodes phonological distinctions specific to Santali that have no standard mapping to either target script. Santali is therefore excluded from all unified-script conditions (Section 3).

Corpus Cleaning and Split Construction We applied Unicode NFC normalisation, script-consistency checks, whitespace cleanup, and punctuation normalisation. Length filtering removed sentences with fewer than 6 or more than 80 words, following thresholds used in BHEPC (Singh et al., 2025a) and FLORES-200 (Costa-Jussà et al., 2022). Near-duplicate removal used cosine similarity at the token level; pairs above 0.95 were collapsed to a single instance to prevent redundancy and split leakage. Script-consistency and length-ratio checks served as review flags rather than automatic rejection rules, since loanwords, numerals, named entities, and code-mixed forms can be valid in these communities. All flagged examples were accepted, corrected, or excluded after review by a fluent speaker. Train, development, and test splits were constructed after deduplication with strict sentence-level disjointness. The split ratio is approximately 90/5/5 per language.

Santali/Ol Chiki: a principled exclusion. Santali is evaluated only in corpus-native Ol Chiki. The script was designed in 1925 specifically to encode Santali phonological distinctions and has be-

come central to Santali education and community identity. Neither Devanagari nor Latin conversion is feasible: no existing transliteration model including IndicXlit supports Ol Chiki→Devanagari or Ol Chiki→Latin conversion, because Ol Chiki encodes phonological contrasts (e.g., checked vs. unchecked vowels, glottal stops) that have no standard character-level mapping in either target script. This is not a limitation of our pipeline but a property of the script itself: Ol Chiki was designed to be non-interconvertible with Devanagari for precisely these phonological reasons.

Santali carries the highest tokenisation premium in the corpus ($2.49\times$ on NLLB), making it the language for which script unification would offer the largest efficiency gain. That it must nonetheless be excluded demonstrates that STP is a necessary but not sufficient criterion: speaker-grounded transliteration validity is the gate.

Corpus-native baselines. The corpus-native script condition provides the baseline for each language: Devanagari for Gondi, Mundari, and Bhili, Odia for Kui, Ol Chiki for Santali, and Latin for Garo and Kokborok. Script-conversion effects are reported as changes relative to this corpus-native baseline. Non-zero effects therefore indicate how a model responds to a particular script-conditioned representation, rather than artifacts of changing the underlying sentences, splits, or references.

D Translation Guidelines and Verification Protocol

Guidelines were developed jointly with senior speakers during a pilot round before full-scale collection, then refined for each language by a designated language coordinator. A shared framework governed semantic fidelity, named-entity treatment, numeral conventions, and cultural-vocabulary preservation across all seven languages. Language-specific supplements addressed orthographic conventions (e.g., aspiration marking in Mundari Devanagari, tone-related spellings in Kokborok Latin), kinship-system vocabulary, and ritual or governance terminology particular to each community.

Semantic fidelity. Translators reproduced meaning, tone, register, and pragmatic intent faithfully - formal, colloquial, or emphatic without addition or deletion. Minor typographical or grammatical

issues in the source could be silently corrected, but factual content was retained exactly. Translations were expected to read naturally under standard community usage rather than mirror Hindi or English word order.

Named entities. Where a community-established rendering existed for a place name, historically significant person, or government institution, that conventional form was used. Otherwise, entities were transliterated following community orthographic norms. Translators were explicitly instructed not to invent alternative renderings or substitute culturally distant equivalents—for instance, replacing a local festival name with a pan-Indian counterpart.

Numerals, units, and dates. Numeric formats followed the source: spelled out where the source spelled out, in digits where the source used digits. Local counting conventions (*lakh*, *crore*) were used where natural; *billion* and *trillion* were retained when those were the source forms. Units of measurement were preserved without conversion.

Cultural and community-specific vocabulary. When the source contained culturally grounded terms—kinship relations, ritual practices, agricultural implements, governance roles translators preserved the community-specific term rather than substituting a generic Hindi or English paraphrase. Each such decision was logged for downstream audit. This guideline directly addresses the cultural-term substitution failure mode (E2) documented in Section 6: if references themselves flatten community vocabulary, the evaluation cannot detect when a model does the same.

Two-linguist verification. Every sentence was independently re-examined by two expert linguists who had not participated in the original translation. Each verifier checked semantic fidelity, script consistency, named-entity treatment, numeral handling, and cultural-vocabulary preservation, and flagged any error. Where both verifiers accepted the translation, it entered the evaluation references; where either flagged an issue, the sentence was returned for revision and re-verified. We use this consensus protocol rather than reporting an inter-annotator agreement statistic because the qualified speaker pool for several languages numbers fewer than ten literate adults, making κ -style coefficients statistically unstable and uninterpretable across communities of very different sizes.

E Annotation and Data Collection Protocol

E.1 Annotator Assignment

Annotators were assigned by verified language proficiency. Bilingual contributors (X + Hindi or X + English) validated the corresponding translation pairs; Hindi–English bilingual annotators performed cross-language consistency checks only and did not evaluate endangered-language content.

E.2 Data Curation Pipeline

Community coordinators in each target region—Madhya Pradesh (Bhili), Chhattisgarh (Gondi), Jharkhand (Mundari), Jharkhand and West Bengal (Santali), interior Odisha (Kui), western Meghalaya (Garo), and Tripura (Kokborok) were engaged to identify contributors with active literacy in the corpus script. Script literacy is not universal even among fluent speakers: Odia-script literacy is uncommon among Kui speakers, and Ol Chiki literacy is unevenly distributed among Santali speakers, requiring additional contributor screening and multiple verification passes for these languages. Pilot rounds of 50–100 seed translations per language were reviewed by senior community members before full-scale collection, and recurring issues (orthographic inconsistencies in Garo and Kokborok Latin, aspiration-marking variation in Mundari Devanagari, Odia spelling conventions for Kui loanwords) were resolved to produce language-specific style guides distributed to all contributors.

E.3 Collection Workflow and Quality Control

Corpus construction ran over three months with approximately five native speakers per language. Contributors translated, typed, reviewed, and corrected sentences in the corpus script. The process was coordinated with language experts, teachers, senior community members, and individuals with experience in language documentation. Contributors were compensated at government daily-wage norms through the corresponding tribal research institutes. Seeds were drawn from public-domain or permissively licensed sources across six domains: agriculture, health, local governance, education, everyday dialogue, and the cultural data. Sentences presupposing institutional or urban contexts without a community equivalent were removed at screening. A second speaker verified every sentence for meaning fidelity, script consistency.

Table 9: Script-consistency audit for corpus-native and converted script conditions. Latin and Devanagari columns report the percentage of sentences containing characters from each script, while consistency reports the proportion of sentences matching the intended condition. The lower consistency for Kokborok in the Devanagari condition reflects residual Latin loanwords and names that were not converted.

Lang.	Cond.	Latin %	Deva %	Consist.
Bhili	native	2.8	100.0	.997
Bhili	latin	100.0	0.0	1.000
Gondi	native	1.4	100.0	.998
Gondi	latin	100.0	0.0	1.000
Mundari	native	6.8	100.0	.988
Mundari	latin	100.0	0.0	1.000
Garó	native	100.0	0.0	1.000
Garó	deva	21.8	100.0	.996
Kokborok	native	100.0	0.0	1.000
Kokborok	deva	95.7	100.0	.957
Kui	native	1.1	65.5	.986
Kui	latin	100.0	0.0	1.000
Kui	deva	1.1	99.0	1.000

tency, cultural-term treatment, and possible source copying; flagged examples were corrected, excluded, or moved out of evaluation splits.

Quality control combined Unicode NFC normalisation, script-consistency checks, length-ratio filtering, and manual review of flagged examples.

E.4 Gold Reference Policy

Every development and test sentence is human-curated, post-edited, or explicitly verified by a fluent speaker. model-generated text is never used as an endangered-language gold reference at any evaluation tier.

F Bhili Lexical Overlap and Leakage Analysis

Bhili shares Devanagari script and substantial lexical material with Hindi, particularly in the governance and education domains represented in our corpus. Under the IndicTrans2 tokeniser, Bhili–Hindi subword vocabulary overlap reaches $ov = 0.608$, the highest in the corpus.

This overlap creates two confounds. First, the Bhili→Hindi chrF++ of 79.2 under IndicTrans2 reflects both valid translation and surface copying: a native-speaker audit found that portions of the high score arise from source tokens passing through unchanged rather than being genuinely translated. This pattern is independently documented in BHEPC (Singh et al., 2025a), which uses 100k training sentences and similarly flags Bhili–Hindi as requiring human verification. All Bhili→Hindi scores are therefore marked with †

throughout the paper and should be interpreted as optimistic upper bounds on genuine cross-lingual transfer quality.

Second, proximity to Hindi in the Devanagari vocabulary cluster positions Bhili close to neighbouring Devanagari-script language varieties, helping explain the E5 drift errors documented in Section 6. The same vocabulary overlap that inflates Bhili→Hindi scores can also encourage target-language drift in the reverse direction.

The model comparison supports this interpretation. On Bhili→Hindi, NLLB falls 36.1 chrF++ below IndicTrans2, whereas ByT5 stays within 1.9 chrF++ (Table 32). Because ByT5 operates on raw bytes rather than a learned subword vocabulary, this near-parity suggests that part of IndicTrans2’s apparent advantage on Bhili comes from vocabulary-level string overlap, not only from deeper cross-lingual transfer.

We retain Bhili rather than excluding it because the confound is itself informative: high lexical overlap can inflate automatic metrics in ways that only speaker verification can detect. This reinforces the need for community-accountable evaluation in endangered-language MT.

G Tokenisation Metrics and Vocabulary Overlap

This appendix formalises the diagnostic metrics introduced in Section 4, reports the tokenisation statistics that underpin the script-effect analysis in Section 5, and examines the relationship between vocabulary overlap and translation quality across the full language–script–tokeniser grid.

G.1 Definitions

Subword Fertility.

$$\text{Fert}(L, S, T) = \frac{\text{total subword tokens}(L, S, T)}{\text{total whitespace words}(L, S)}. \quad (4)$$

Sentence-level fragmentation $\text{SFrag}(L, S, T)$ is the mean subword token count per sentence for language L , script condition S , and tokeniser T . Higher fertility produces longer sequences, weaker per-token information density, and harder cross-lingual parameter reuse (Rust et al., 2021).

Script Tokenisation Parity.

$$\text{STP}(L, S, T) = \frac{\text{SFrag}(L, S, T)}{\text{SFrag}(L, S_{\text{nat}}, T)}. \quad (5)$$

Here S may be a unified-script condition such as Devanagari or Latin. $STP < 1$ indicates reduced fragmentation relative to the corpus-native script, $STP > 1$ indicates increased fragmentation, and $STP \approx 1$ is expected when the evaluated script condition is identical to the corpus-native script condition.

Tokenisation Premium.

$$\text{Prem}(L, S, T) = \frac{\text{SFrag}(L, S, T)}{\text{SFrag}(r, S_{\text{nat}}, T)}, \quad (6)$$

$$r \in \{\text{Hi}, \text{En}\}.$$

This measures how many more subword tokens a language requires relative to a high-resource reference under the same tokeniser. Cross-checked values from Petrov et al. (2023): Santali (Ol Chiki) $2.49\times$ vs. English on NLLB, $12.86\times$ on GPT-2; Hindi (Devanagari) $1.22\times$ on NLLB.

Vocabulary Overlap.

$$\text{Ov}(L_i, L_j, S, T) = \frac{|V_i \cap V_j|}{|V_i \cup V_j|}, \quad (7)$$

$$L_j \in \{\text{Hi}, \text{En}\}.$$

where $V_i = V(L_i, S, T)$ denotes the set of subword types observed for language L_i under script S and tokeniser T .

Fragmentation change and MT change.

$$\Delta\text{Frag} = \text{Fert}_{L, S_{\text{nat}}, T} - \text{Fert}_{L, S_{\text{uni}}, T},$$

$$\Delta_{\text{MT}} = \text{chrF}^{++}_{L, S_{\text{uni}}, M} - \text{chrF}^{++}_{L, S_{\text{nat}}, M}. \quad (8)$$

Positive ΔFrag means unification reduces fragmentation; positive Δ_{MT} means translation improves. The fragmentation-mediated transfer hypothesis predicts a positive association between reduced fertility and translation gain, but deviations are expected when vocabulary overlap or subword quality changes independently of sequence length.

G.2 Tokenisation Statistics

Table 10 reports fertility, STP, tokenisation premium, and Hindi vocabulary overlap for representative conditions, and Table 36 extends fertility, average sequence length, STP, and Hindi overlap to all 57 language-script-tokenizer conditions. Table 13 additionally reports multi-piece rates and observed vocabulary utilisation for Kui, the only language with three genuinely distinct representations. The two STP values that anchor the

main-body reversal analysis are Kui Devanagari under IT2 ($STP = 0.396$, the largest fragmentation reduction in the corpus) and Garo Devanagari under NLLB ($STP = 1.619$, the largest increase). Neither extreme predicts its own outcome. The largest fragmentation reduction in the corpus, Kui Devanagari under IT2, yields $\Delta_{\text{MT}} = -1.8$, whereas the same representation gives $+19.3$ under NLLB; conversely the largest fragmentation increase, Garo Devanagari under NLLB, still improves translation by $+12.0$ chrF++, and Garo Devanagari also gains $+6.7$ under IT2 (Table 3). Neither the direction nor the magnitude of the STP change therefore determines the direction of Δ_{MT} , confirming that STP is a necessary but not sufficient predictor of translation quality change. Figure 6 provides a complementary aggregate view: panels (a)–(b) show that ByT5’s byte-level encoding produces dramatically higher fertility (12–18 subwords per word) than either subword tokeniser (2–4), while panels (c)–(d) reveal that the inter-language variance in STP is largest under IT2, consistent with its Devanagari-concentrated vocabulary creating the widest spread between well-covered and poorly covered scripts.

G.3 Vocabulary Overlap and Its Consequences

Table 11 reports Hindi overlap at two granularities: word-level Jaccard over whitespace-tokenised types, and subword-level Jaccard under each of the three tokenisers. The overlap ordering (Bhili $\gg$ Mundari $>$ Gondi $\gg$ Garo $\approx$ Kui $\approx$ Kokborok $\approx$ Santali) is stable across both measurement granularities and across all three tokenisers.

Three patterns in this table connect directly to the main-body findings. First, ByT5 shows uniformly high overlap (>0.69) because UTF-8 byte sequences are shared across scripts far more broadly than learned subword units, this is the vocabulary-level explanation for ByT5’s low script sensitivity (Table 4). Second, the sharp drop from Devanagari-script languages (Bhili, Mundari, Gondi: 0.44 – 0.61 under IT2) to non-Devanagari languages (Garo, Kui, Kokborok, Santali: <0.04 under IT2) explains why IndicTrans2 produces strong results for the former group and severe generation failures for the latter, the tokeniser’s Devanagari-heavy vocabulary simply does not encode the subword units these languages need. Third, Bhili’s exceptional overlap with Hindi

Table 10: Tokenization metrics for selected script conditions. Fertility reports subword tokens per whitespace word; STP measures fragmentation relative to the corpus-native condition; $Prem_{En}$ and $Prem_{Hi}$ compare fertility with English and Hindi references; and ov_{Hi} reports subword Jaccard overlap with Hindi.

Language / Script	Tok.	Fert	STP	$Prem_{En}$	$Prem_{Hi}$	ov_{Hi}
Santali – Ol Chiki (native)	NLLB	2.85	1.00	2.49	2.04	0.060
Kui – Odia (native)	IT2	6.05	1.00	5.26	5.50	0.010
Kui – Devanagari	IT2	2.39	0.396	2.08	2.18	0.207
Kui – Latin	IT2	4.06	0.671	3.53	3.69	0.051
Kui – Odia (native)	NLLB	3.12	1.00	2.73	2.24	0.044
Garó – Latin (native)	NLLB	2.56	1.00	2.24	1.83	0.036
Garó – Devanagari	NLLB	4.14	1.619	3.62	2.97	0.213
Bhili – Devanagari (native)	IT2	1.38	1.00	1.20	1.26	0.608
Gondi – Devanagari (native)	IT2	1.83	1.00	1.59	1.66	0.437
Mundari – Devanagari (native)	IT2	2.09	1.00	1.82	1.90	0.574
Kokborok – Latin (native)	IT2	2.63	1.00	2.28	2.39	0.032
Hindi – Devanagari (ref.)	NLLB	1.39	–	1.22	1.00	1.000

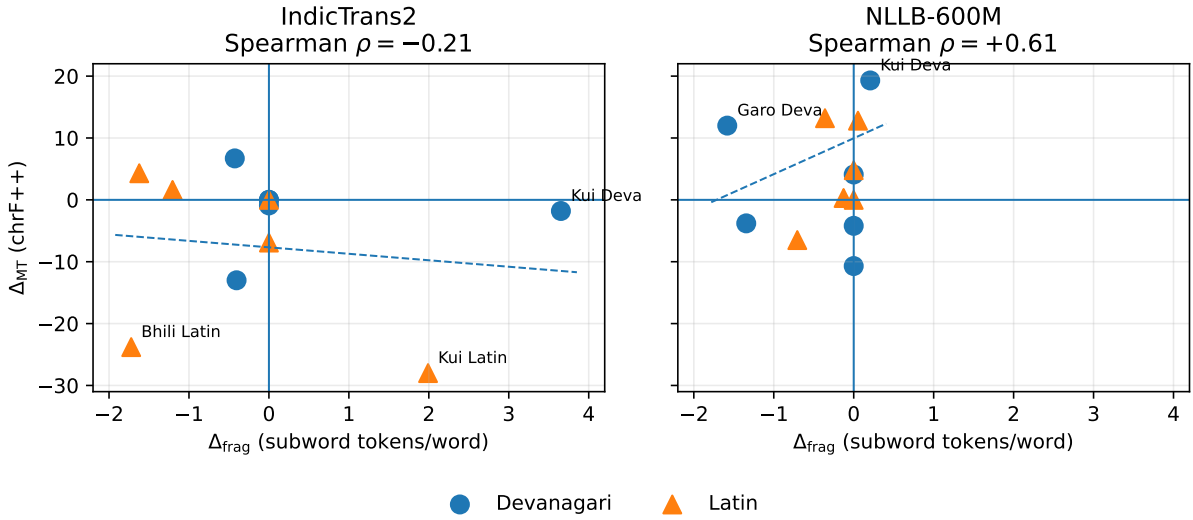


Figure 5: Fragmentation change and translation change in X→Hindi. Reduced fertility shows a positive association with chrF++ gain under NLLB-600M ($\rho = +0.61$), but not under IndicTrans2 ($\rho = -0.21$), confirming that shorter tokenization alone does not guarantee better transfer.

(0.608 under IT2, 0.781 under NLLB) quantifies the copying confound documented in Appendix F: a tokenizer that encodes Bhili and Hindi with largely the same subword units cannot reliably distinguish translation from surface copying, inflating automatic scores for the Bhili–Hindi pair.

H MT Systems, Training, & Compute Details

Section 4 describes the three model families and their role in the experimental design. This appendix reports training configuration and compute details for reproducibility.

Training. All fine-tuning ran on NVIDIA A100 GPUs using HuggingFace Transformers. We

used the batch sizes 16 and learning rates $\{10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$ with AdamW, applying the same grid uniformly across all languages, directions, and script conditions to prevent any representation from receiving a tuning advantage. Checkpoints were selected by best development chrF++ with early stopping at patience 5. Decoding used deterministic beam search (beam 5) with maximum generation length $\min(2 \times |src|, 256)$ tokens. Training time ranged from 6 to 24 GPU-hours per run depending on model size. We fine-tune IndicTrans2-320M and NLLB-600M as contrasting subword-based multilingual MT systems, and ByT5-small, a 300M-parameter byte-level UTF-8 model, as a diagnostic control for removing learned vocabulary allocation. For the ca-

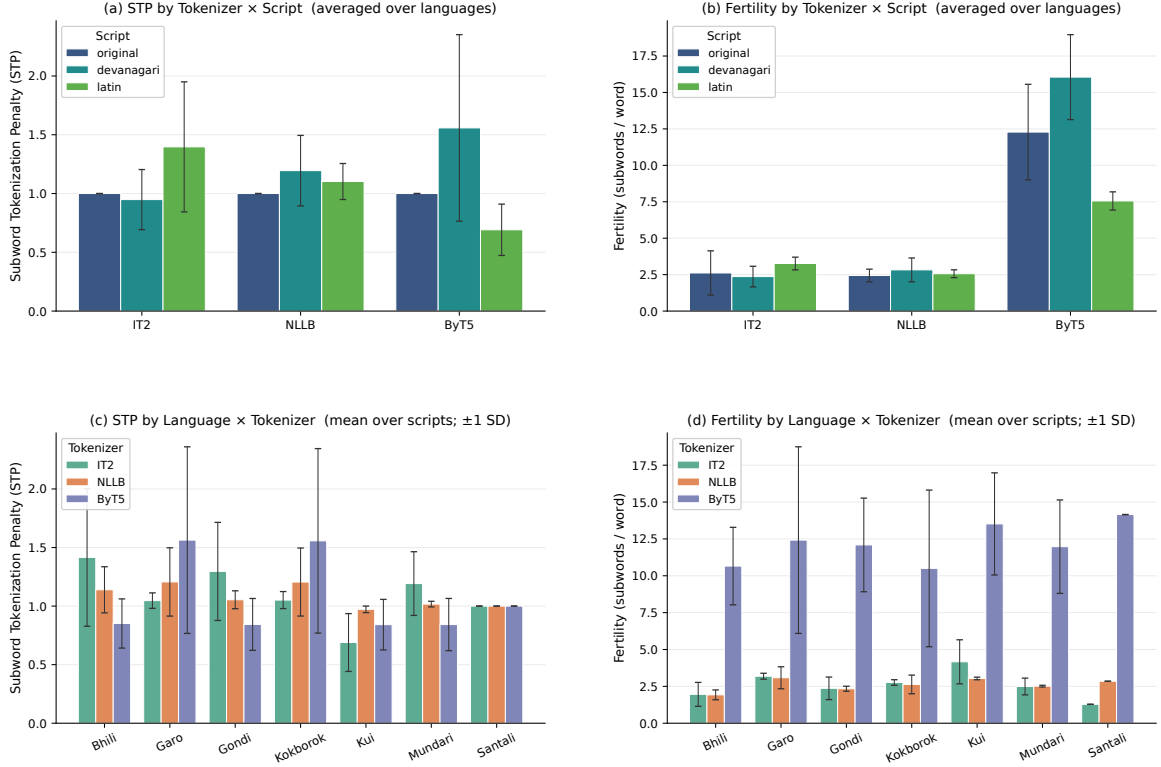


Figure 6: Tokenisation efficiency by tokenizer, script condition, and language. Panels (a)–(b) summarize average STP and fertility by tokenizer–script condition; panels (c)–(d) show language-level variation. Error bars denote ± 1 standard deviation, showing that script effects are tokenizer- and language-dependent.

Table 11: Vocabulary overlap with Hindi at word and subword levels. Jaccard similarity is computed over whitespace-tokenised word types and tokenizer-specific subword types. The table contextualizes transfer and possible score inflation by showing that Bhili, Mundari, and Gondi share far more surface material with Hindi than Garo, Kui, Kokborok, and Santali.

Language–Hindi	Word		Subword	
	Jaccard	IT2	NLLB	ByT5
Bhili	0.268	0.608	0.781	0.984
Mundari	0.181	0.574	0.767	0.937
Gondi	0.122	0.437	0.630	0.895
Garo	0.010	0.036	0.036	0.782
Kui	0.001	0.010	0.044	0.728
Kokborok	<0.001	0.032	0.036	0.692
Santali	<0.001	0.024	0.060	0.864

capacity check reported in Appendix J (Table 26), we additionally fine-tune ByT5-base (~ 580 M parameters) on Kui using the same training and decoding settings across all three script conditions.

Prompted baselines. We evaluated GPT-4o-mini, Llama 3 8B-Instruct, and Mistral 7B-Instruct in zero-shot and five-shot settings. The five-shot examples were sampled from the training split,

Table 12: Relationship between fragmentation change (ΔFrag) and MT change (ΔMT) across all four translation directions. Results pool the seven valid script conversions, giving 28 observations per model. Spearman’s ρ shows a stronger relationship between reduced fragmentation and MT gain for NLLB-600M than for IndicTrans2, consistent with the model-dependent pattern in Figure 5. ByT5-small operates over byte sequences, so its slope is not directly comparable to those of the subword models.

Model	Spearman ρ	Slope	95% CI
IndicTrans2	+0.14	−0.4	[−4.1, +15.4]
NLLB-600M	+0.56	+5.9	[−2.2, +16.7]
ByT5-small	+0.35	+0.25	[+0.11, +0.41]

with no selection on the development set. We used the same prompt template for all languages and translation directions (Appendix I.1).

I Prompt Templates and Prompting Results

I.1 Templates

Zero-shot prompting uses corpus-native script only. Five-shot prompting is evaluated under all valid script conditions (corpus-native, Devanagari-

Table 13: Tokenization statistics and MT change for Kui across script representations under IndicTrans2 and NLLB-600M. Alongside fertility and STP, we report the average number of subword pieces per sentence, the proportion of words split into three or more pieces, observed vocabulary utilisation, and the corresponding change in chrF++ relative to the corpus-native condition (Δ_{MT}). Vocabulary utilisation is tokenizer-specific and is not directly comparable across models.

Model	Repr.	Fert./STP	Pcs/sent.	3+ pcs	Vocab util.	Δ_{MT}
IT2	Odia (nat.)	6.05 / 1.000	71.9	92.3%	323 (0.26%)	0.0
IT2	Devanagari	2.39 / 0.396	28.5	40.9%	11,448 (9.33%)	-1.8
IT2	Latin	4.06 / 0.671	48.2	76.2%	2,515 (2.05%)	-28.0
NLLB	Odia (nat.)	3.12 / 1.000	37.1	62.9%	1,975 (0.77%)	0.0
NLLB	Devanagari	2.92 / 0.934	34.6	57.6%	3,796 (1.48%)	+19.3
NLLB	Latin	3.07 / 0.983	36.5	59.6%	7,624 (2.98%)	+12.8

unified, Latin-unified); Santali is always Ol Chiki only. For X-target directions, expected output script follows the evaluated condition; Hindi and English targets use Devanagari & Latin respectively.

Zero-shot.

Translate from [SRC_LANG] to [TGT_LANG].
[SRC_LANG]: [SOURCE]
[TGT_LANG]:

Five-shot.

Translate from [SRC_LANG] to [TGT_LANG]. Produce only the translation.
[SRC_LANG]: [SRC_1] [TGT_LANG]: [TGT_1]
... (five pairs from training split) ...
Now translate:
[SRC_LANG]: [SOURCE]
[TGT_LANG]:

I.2 Corpus-Native Prompting Results

I.3 Unified-Script Prompting

Devanagari-unified prompting for Garo and Kokborok frequently shifts the output language rather than improving translation quality, illustrating the language-routing instability that the tokeniser-mediated account predicts: when the source representation falls outside the model’s vocabulary coverage, the decoder defaults to the nearest high-resource neighbour rather than producing the requested target. Latin-unified prompting on Bhili and Kui produces quality drops consistent with the fine-tuned results in Table 3, confirming that the script-tokeniser interaction is visible even before task-specific adaptation.

J Fine-Tuned Results: All Script Conditions

Tables 27 and 28 report fine-tuned results (BLEU/chrF++) for the Devanagari-unified and Latin-unified conditions; corpus-native results appear in Table 5. Santali is excluded from unified-script conditions (Section 3).

Figure 7 provides a visual summary of the corpus-native results across all four directions. The colour gradients make two patterns immediately visible: Santali and Bhili occupy the warm end of the scale under IndicTrans2 (reflecting pretraining coverage and lexical overlap respectively), while Kui occupies the cold end in the En→X direction (reflecting decoder-side vocabulary depletion). The directional contrast X→Hindi and X→English are systematically warmer than Hi→X and En→X reflects the asymmetry between well-resourced target-side generation and sparse source-side encoding documented in Section 6. The $6.8\times$ IT2/ByT5 ratio confirms that the learned subword vocabulary drives most observed script sensitivity. Residual ByT5 sensitivity (≈ 2 chrF++) reflects a byte-length asymmetry: Devanagari and Odia characters occupy 3 UTF-8 bytes versus 1–2 for Latin, producing a sequence-length difference that persists without a learned vocabulary.

K Statistical Analysis of Tokenizer-Mediated Effects

K.1 Paired Bootstrap Evaluation Protocol

All pairwise model comparisons reported use paired bootstrap resampling (Koehn, 2004) with $N=1,000$ resampling iterations over matched sentence-level chrF++ scores. In every comparison the reference system is IndicTrans2 fine-tuned on the corpus-native script condition for the same language and direction, so the test statistic is

$$\Delta chrF++ = chrF++_{\text{model}} - chrF++_{\text{IT2-native}}. \quad (9)$$

We report two-sided p -values and 95% bootstrap confidence intervals (CI). A difference is considered significant when $p<0.05$ and the CI excludes zero; the stricter threshold $p<0.001$ is also noted throughout. Where two comparisons share the same language–direction stratum (NLLB-600M vs. IT2 and ByT5 vs. IT2), we additionally apply the Bonferroni correction at $\alpha=0.05$, yielding

Table 14: Corpus-native prompting results for Hindi→X translation. Each cell reports BLEU/chrF++ for zero-shot and five-shot settings across the seven target languages. The table provides the prompting baseline used to measure the adaptation gains from fine-tuned MT systems.

Model	Shot	Gondi	Mundari	Bhili	Santali	Kui	Garó	Kokborok
GPT-4o-mini	0	1.30/18.34	3.42/15.29	5.89/35.65	0.05/0.90	0.22/1.38	1.76/21.39	1.26/18.34
GPT-4o-mini	5	4.03/21.35	6.43/23.99	9.84/39.43	0.74/10.70	1.05/14.01	4.34/27.61	3.08/26.93
Llama 3 8B	0	0.66/13.61	2.90/18.09	4.62/27.85	0.03/0.68	0.22/0.96	0.13/2.34	0.15/1.51
Llama 3 8B	5	0.74/16.11	2.47/17.31	12.72/31.76	1.63/9.32	1.00/7.36	2.19/16.92	3.01/19.62
Mistral 7B	0	0.69/4.35	1.70/9.41	6.95/20.08	0.01/0.53	0.01/0.24	0.05/0.91	0.03/0.42
Mistral 7B	5	2.53/12.12	4.17/16.56	9.79/27.13	0.54/6.58	0.12/4.50	1.95/16.92	1.99/17.47

Table 15: Corpus-native prompting results for X→Hindi translation. Each cell reports BLEU/chrF++ for zero-shot and five-shot prompting across the seven target languages. This table provides the prompt-only baseline for measuring how much task-specific fine-tuning improves translation into the high-resource Hindi target.

Model	Shot	Gondi	Mundari	Bhili	Santali	Kui	Garó	Kokborok
GPT-4o-mini	0	7.27/25.80	8.68/26.95	29.75/52.41	0.26/14.00	0.62/13.77	6.37/28.15	4.34/23.29
GPT-4o-mini	5	9.40/28.63	11.05/29.60	34.18/56.01	3.35/17.08	1.96/15.51	9.55/31.68	7.28/27.11
Llama 3 8B	0	2.42/19.57	4.35/21.05	11.11/34.32	0.47/12.47	0.29/10.81	5.93/20.24	4.33/16.48
Llama 3 8B	5	2.89/20.31	4.94/22.59	18.66/36.87	0.61/13.78	1.14/11.79	8.00/23.33	6.56/20.33
Mistral 7B	0	3.72/16.82	5.50/18.36	11.65/27.31	0.11/7.59	1.46/10.84	2.83/15.53	2.49/12.81
Mistral 7B	5	3.50/17.17	6.65/19.83	11.95/27.92	0.14/10.69	1.54/11.37	5.33/19.66	5.24/18.19

Table 16: Corpus-native prompting results for English→X translation. Each cell reports BLEU/chrF++ for zero-shot and five-shot prompting across the seven target languages. This table evaluates whether general-purpose LLMs can generate the target languages without fine-tuning, the hardest setting for languages with limited pretraining support.

Model	Shot	Gondi	Mundari	Bhili	Santali	Kui	Garó	Kokborok
GPT-4o-mini	0	0.31/10.14	0.91/1.56	0.63/21.54	0.04/0.16	0.18/0.36	3.48/25.30	3.87/25.73
GPT-4o-mini	5	2.01/14.82	3.67/17.92	2.07/23.57	0.48/10.17	0.33/13.40	6.30/30.29	6.07/29.33
Llama 3 8B	0	0.15/9.14	1.00/8.99	0.24/11.38	0.03/0.48	0.47/5.48	0.79/6.13	0.69/1.74
Llama 3 8B	5	0.13/9.66	0.99/11.85	1.73/14.72	1.41/8.89	0.50/6.51	5.07/21.07	6.54/23.67
Mistral 7B	0	0.15/8.63	1.29/9.72	0.80/12.02	0.01/0.45	0.03/0.38	2.00/3.96	0.92/2.47
Mistral 7B	5	0.27/10.10	2.14/12.12	1.17/13.65	0.83/6.87	0.19/4.43	5.73/23.80	5.23/22.20

Table 17: Corpus-native prompting results for X→English translation. Each cell reports BLEU/chrF++ for zero-shot and five-shot prompting across the seven target languages. This table measures prompt-only performance when the model encodes the target-language source but decodes into high-resource English.

Model	Shot	Gondi	Mundari	Bhili	Santali	Kui	Garó	Kokborok
GPT-4o-mini	0	2.87/25.04	6.25/30.30	8.81/36.46	0.48/19.93	2.39/21.04	10.83/35.47	7.44/31.33
GPT-4o-mini	5	5.13/25.35	9.02/32.59	11.13/38.82	2.16/21.15	2.88/20.76	13.15/38.27	9.88/33.58
Llama 3 8B	0	1.48/20.39	2.52/23.83	3.19/26.51	0.79/19.51	1.27/16.43	13.40/32.03	9.06/27.09
Llama 3 8B	5	1.06/19.52	3.30/24.73	4.71/27.89	1.30/17.80	1.02/16.59	14.36/33.61	8.77/27.25
Mistral 7B	0	1.05/17.58	5.20/22.43	3.31/23.26	0.21/14.49	0.91/14.92	14.91/31.72	11.67/27.29
Mistral 7B	5	0.80/19.65	4.83/22.89	3.85/25.53	0.84/18.31	0.93/16.55	16.06/34.05	12.55/28.90

a per-comparison threshold of $p < 0.025$; all conclusions below are unchanged under this stricter criterion.

K.2 Aggregate Results Across Translation Directions

Tables 31–34 summarize all 56 pairwise comparisons. Of these, 52 are significant at

Table 18: Five-shot prompting results for Hindi→X under the Devanagari-unified script condition. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Devanagari was not supported; see Section 3.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	2.56/18.99	4.87/23.71	5.35/25.43	1.69/13.98	0.99/15.58	0.96/14.86
Llama 3 8B	3.23/15.74	6.12/18.46	11.10/31.16	0.77/9.29	1.72/11.15	3.11/12.75
Mistral 7B	1.78/11.72	4.69/17.21	8.72/26.09	0.40/4.60	0.22/2.40	0.45/1.99

Table 19: Five-shot prompting results for X→Hindi under the Devanagari-unified script condition. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Devanagari was not supported; see Section 3.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	8.45/26.70	9.88/30.43	28.45/51.54	3.86/17.67	5.03/24.63	3.88/20.69
Llama 3 8B	5.27/20.55	7.81/22.86	18.33/37.78	1.79/13.04	2.11/15.75	1.88/14.72
Mistral 7B	4.33/18.11	5.96/18.85	11.01/27.56	1.54/12.29	1.67/14.32	2.18/14.85

Table 20: Five-shot prompting results under the Devanagari-unified script condition for English→X translation. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Devanagari was not supported.; see Section 3.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	1.03/13.57	2.54/18.44	3.61/20.61	0.53/12.91	1.85/17.61	1.64/16.45
Llama 3 8B	1.48/10.20	2.71/11.92	1.20/13.24	0.52/7.48	3.07/13.34	3.99/14.87
Mistral 7B	1.05/10.78	2.36/12.37	0.80/13.40	0.49/7.77	0.81/6.36	1.54/6.91

Table 21: Five-shot prompting results under the Devanagari-unified script condition for X→English translation. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Devanagari was not supported.; see Section 3.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	3.92/25.07	8.07/32.49	1.89/15.96	3.56/23.03	8.14/31.45	5.75/26.64
Llama 3 8B	1.76/19.65	6.17/24.69	4.36/28.10	1.42/17.39	2.86/22.35	2.10/20.54
Mistral 7B	1.70/19.41	5.43/23.14	3.76/25.67	2.34/17.53	3.27/23.02	4.37/22.24

Table 22: Five-shot prompting results under the Latin-unified script condition for Hindi→X translation. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Latin was not supported.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	1.91/19.15	3.77/24.38	4.19/34.93	0.29/14.69	4.71/26.79	3.40/26.82
Llama 3 8B	4.00/16.62	5.69/17.79	10.17/25.09	1.41/10.99	1.62/16.78	2.84/20.35
Mistral 7B	4.43/16.85	6.11/18.88	9.51/25.61	1.69/11.42	1.23/16.17	2.21/17.74

Table 23: Five-shot prompting results under the Latin-unified script condition for X→Hindi translation. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Latin was not supported.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	6.10/23.13	7.98/28.15	3.41/19.71	3.49/16.68	10.22/32.04	8.18/27.38
Llama 3 8B	2.39/16.17	4.45/19.08	9.76/29.15	1.09/12.20	7.57/22.69	6.59/20.39
Mistral 7B	2.64/17.11	4.93/18.59	7.86/25.87	1.49/12.52	4.54/18.31	4.84/17.96

$p < 0.001$ and 54 at $p < 0.05$. The two non-significant comparisons are Garó→Hindi under ByT5 ($\Delta = +0.06$, 95% CI: $[-0.99, +1.00]$, $p = 0.896$) and Mundari→English under ByT5 ($\Delta = -0.73$, 95% CI: $[-2.00, +0.61]$, $p = 0.298$). Both are discussed in §K.4.

Table 24: Five-shot prompting results under the Latin-unified script condition for English→X translation. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Latin was not supported.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	1.22/12.93	2.35/16.13	0.49/20.78	0.50/13.38	8.50/30.21	6.50/29.75
Llama 3 8B	1.74/10.88	2.03/11.52	0.94/10.77	0.67/8.96	4.77/21.53	6.10/22.37
Mistral 7B	1.82/10.73	0.99/11.16	0.53/7.83	0.51/8.92	4.96/22.72	4.67/22.42

Table 25: Five-shot prompting results under the Latin-unified script condition for X→English translation. Each cell reports BLEU/chrF++. Santali is excluded because lossless Ol Chiki conversion to Latin was not supported.

Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
GPT-4o-mini	3.14/22.11	5.90/29.06	7.37/33.06	2.99/22.17	15.81/40.13	9.79/33.50
Llama 3 8B	1.15/16.45	3.09/19.45	1.36/19.66	1.09/16.30	12.58/32.40	8.91/28.47
Mistral 7B	0.98/17.80	3.70/20.96	1.18/22.02	1.12/17.58	15.09/33.00	11.42/28.92

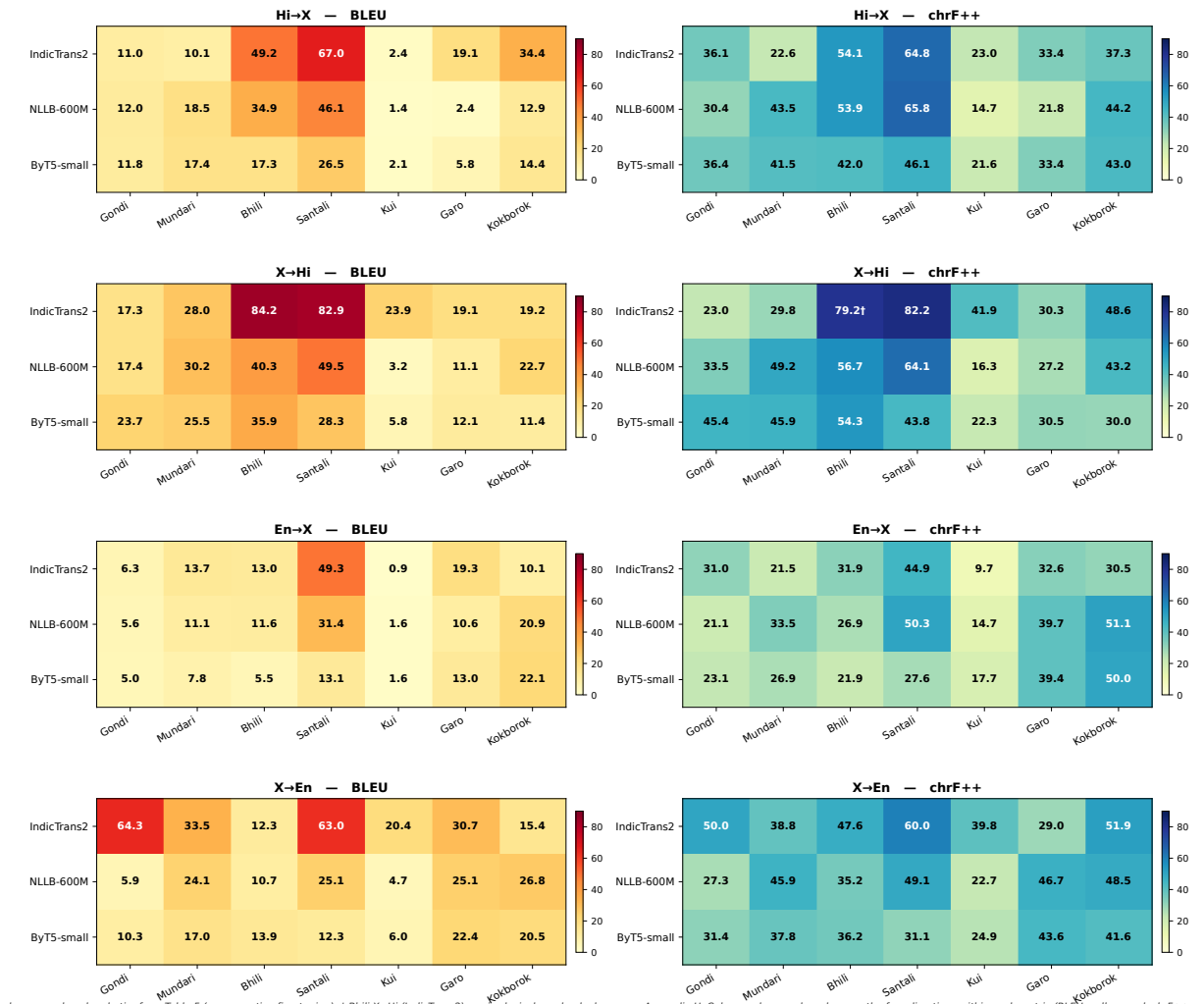


Figure 7: Corpus-native fine-tuning results by model, language, and direction. Left: BLEU; right: chrF++. Values correspond to Table 5. [†]Bhili X→Hi (IT2) flagged for lexical-overlap leakage (Appendix F).

K.3 Sign Reversals as Evidence of Tokenizer Dependence

Across the 28 language-direction pairs where both NLLB and ByT5 are evaluated against Indic-

Trans2, 14 exhibit a verified sign reversal: NLLB and ByT5 shift relative to IndicTrans2 in opposite directions, with both differences significant at $p < 0.05$. Two additional pairs (Garó→Hindi and Mundari→English) show nominal sign differ-

Table 26: ByT5-base results for Kui under corpus-native Odia, Devanagari, and Latin script conditions. For Kui→Hindi, chrF++ varies by only 0.45 across scripts despite substantial differences in byte-sequence length. In the reverse Hindi→Kui direction, the Latin representation yields the highest score, showing that reduced source-side script sensitivity does not imply target-side script invariance.

Direction	Script	BLEU	chrF++	Fert.	Byte util.
Kui→Hi	Odia	8.67	25.64	16.06	46.48%
Kui→Hi	Deva	8.57	26.09	15.87	46.09%
Kui→Hi	Latin	8.33	25.67	8.62	48.05%
Hi→Kui	Odia	2.32	22.71	16.06	46.48%
Hi→Kui	Deva	2.40	22.85	15.87	46.09%
Hi→Kui	Latin	2.62	27.42	8.62	48.05%

ences but are excluded from this count because one arm is non-significant; they are discussed in §K.4.

These reversals are difficult to explain through a simple “model A is stronger than model B” narrative. If one architecture were uniformly superior, the signs would remain stable across languages and directions. Instead, the observed reversals indicate that translation quality depends strongly on how tokenizers encode script, morphology, and lexical structure for individual languages. Importantly, the 14 verified reversals span five of the seven languages in the corpus (Gondi, Kui, Garo, Kokborok, and Santali) and appear across all four translation directions, making a direction-specific confound unlikely. The clearest reversals occur for **Gondi**. Across all directions, NLLB falls 7.97–22.08 chrF++ below IndicTrans2 (all $p < 0.001$), whereas ByT5 improves over IndicTrans2 by 1.66–13.67 chrF++ (all $p \leq 0.002$). The confidence intervals are consistently narrow (± 0.5 – 1.0 chrF++), indicating stable effects rather than sampling noise. This behavior suggests that IndicTrans2’s Devanagari-focused SentencePiece vocabulary provides denser subword coverage for Gondi than the vocabularies used by NLLB, while ByT5 partially compensates through byte-level encoding.

Santali shows the most consistent directional pattern in the corpus: NLLB exceeds IT2 in all four directions ($\Delta = +2.08$ to $+13.33$, all $p < 0.001$), while ByT5 falls below IT2 in all four ($\Delta = -12.82$ to -21.16 , all $p < 0.001$). This is the inverse of the Gondi pattern. Santali is the only language in the corpus explicitly represented

in IT2 pretraining as one of the 22 scheduled Indian languages, giving IT2 explicit Ol Chiki subword coverage. NLLB’s broader multilingual vocabulary appears to recover part of this advantage, while ByT5 lacks access to script-specific lexical units altogether. The Santali ByT5 confidence intervals (± 0.5 – 0.9) are among the narrowest in the appendix, suggesting that these differences are highly stable across resampling runs. Together, these reversal patterns indicate that tokenizer design is not merely an implementation detail for endangered-language MT. It directly shapes which scripts and linguistic communities multilingual systems can represent effectively.

K.4 Linguistic Interpretation of Non-Significant Effects

Garo→Hindi, ByT5 ($\Delta = +0.06$, 95% CI: $[-0.99, +1.00]$, $p = 0.896$). ByT5 and IT2 are statistically indistinguishable on this direction despite NLLB falling 21.6 chrF++ below IT2. Garo is a Tibeto-Burman language written in Latin script. Latin characters occupy one to two UTF-8 bytes, so ByT5’s raw byte encoding of Garo input is efficient regardless of any learned vocabulary. This result aligns with the byte-length asymmetry hypothesis: the tokenizer disadvantages affecting Devanagari- and Odia-script languages, where characters typically expand into three-byte sequences are largely absent for Latin-script input.

Mundari→English, ByT5 ($\Delta = -0.73$, 95% CI: $[-2.00, +0.61]$, $p = 0.298$). ByT5 and IT2 perform equivalently on this direction despite NLLB exceeding IT2 by 9.4 chrF++. Mundari is an Austroasiatic agglutinative language; its Devanagari-script representation is longer and more fragmented under both tokenizers than Latin-script languages. On the X→English direction, however, the English decoder is strong in both IT2 and ByT5, and a high-quality generation model can absorb encoding-side variance. The null result therefore does not imply that ByT5 and IndicTrans2 behave similarly on Mundari overall. Indeed, ByT5 significantly outperforms IndicTrans2 on three other Mundari directions (Hi→X: $+10.46$; X→Hi: $+14.39$; En→X: $+4.94$; all $p < 0.001$). Instead, the result suggests that for this particular direction, decoder-side English fluency partially masks tokenization differences on the source side.

Both null results are theoretically coherent:

Table 27: Fine-tuned results under the Devanagari-unified script condition. Each cell reports BLEU/chrF++ across four translation directions. Santali is excluded because lossless Ol Chiki conversion to Devanagari was not supported.

Dir.	Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
Hi→X	IT2	11.04/36.13	11.21/17.91	49.23/54.06	2.17/1.92	10.68/35.35	24.81/33.72
	NLLB	18.97/37.53	19.69/38.55	23.33/41.52	9.83/27.64	12.26/31.79	11.63/41.79
	ByT5	12.62/37.29	16.91/40.57	18.05/42.65	1.69/20.05	5.56/29.54	11.16/36.53
X→Hi	IT2	17.29/23.00	18.44/28.86	84.24/79.15 [†]	32.77/40.07	19.64/37.00	14.22/35.60
	NLLB	21.97/37.53	22.69/38.55	35.33/52.52	17.83/35.64	19.82/39.20	19.35/39.46
	ByT5	23.23/44.90	24.92/44.55	34.63/53.29	4.60/20.41	10.80/28.91	11.05/29.61
En→X	IT2	6.27/30.98	12.30/25.64	12.16/33.03	1.98/1.28	14.54/40.32	19.52/42.26
	NLLB	14.85/32.01	14.13/28.97	19.48/36.36	10.39/29.56	15.40/34.59	18.46/48.94
	ByT5	4.55/22.90	7.73/26.06	5.56/22.21	0.88/15.88	9.46/34.99	16.38/40.52
X→En	IT2	14.06/25.85	32.38/39.15	12.30/47.63	41.97/49.41	42.73/60.67	19.29/51.65
	NLLB	20.11/42.29	19.66/38.63	26.81/38.63	21.39/40.56	24.69/44.13	22.94/44.47
	ByT5	10.32/31.18	16.87/37.55	13.71/35.61	4.94/23.51	19.43/41.58	17.93/39.51

Table 28: Fine-tuned results under the Latin-unified script condition. Each cell reports BLEU/chrF++ across four translation directions. Santali is excluded because lossless Ol Chiki conversion to Latin was not supported.

Dir.	Model	Gondi	Mundari	Bhili	Kui	Garó	Kokborok
Hi→X	IT2	9.86/27.81	5.30/25.38	9.84/38.80	3.67/26.91	19.13/33.43	14.98/38.99
	NLLB	10.68/41.16	19.90/43.69	23.88/49.28	7.69/22.60	9.47/35.01	13.09/43.75
	ByT5	13.59/42.21	20.27/48.42	22.09/51.92	1.69/24.07	6.65/33.83	14.32/43.38
X→Hi	IT2	17.29/27.28	13.61/31.32	56.38/55.36 [†]	8.90/13.93	17.29/23.38	19.18/48.63
	NLLB	22.66/46.70	25.21/49.54	29.30/50.15	13.77/29.16	15.31/32.07	23.37/43.26
	ByT5	21.52/42.57	24.34/43.88	32.77/51.91	3.67/19.15	11.31/29.71	11.78/30.25
En→X	IT2	11.48/26.65	9.84/31.86	11.84/26.08	4.55/26.58	19.30/32.60	10.07/30.54
	NLLB	11.47/26.19	10.92/30.75	15.05/29.75	11.53/25.19	15.50/34.30	21.15/51.42
	ByT5	4.44/26.70	8.99/32.28	5.92/26.63	0.97/16.64	13.36/39.99	22.27/50.34
X→En	IT2	25.41/27.02	49.62/68.61	12.45/31.67	7.99/24.13	41.11/69.08	15.41/51.90
	NLLB	17.08/34.60	16.20/31.60	20.72/39.93	15.53/30.19	20.08/39.10	27.48/48.31
	ByT5	9.31/29.42	15.88/36.36	11.10/32.60	4.27/22.64	23.58/44.87	20.15/41.75

Table 29: Aggregate script sensitivity in X→Hindi translation. Mean and maximum ScSens summarize how strongly each model’s chrF++ varies across valid script conditions. The sharp drop from subword models to ByT5 supports the claim that most cross-script variance is mediated by learned vocabulary allocation.

Model	Mean ScSens	Max ScSens
IndicTrans2	14.2	28.0
NLLB-600M	11.0	19.3
ByT5-small	2.1	3.1

Table 30: Normalised script sensitivity for X→Hindi. Normalising mean ScSens by mean chrF++ over the same six-language set shows that ByT5’s lower script sensitivity is not explained by its score range: its normalised ScSens remains approximately $6.5\times$ lower than that of IndicTrans2. Moreover, Kui—the lowest-scoring case under ByT5—has the highest ScSens among the ByT5 languages.

Model	Mean chrF++	Mean ScSens	Norm. ScSens
IndicTrans2	38.69	14.19	36.7%
NLLB-600M	40.00	10.98	27.4%
ByT5-small	37.07	2.08	5.6%

L Error Annotation Schema and Verification Protocol

they occur precisely where the tokenizer-mediated mechanism predicts reduced or absent effects, and they therefore strengthen rather than weaken the overall mechanistic account.

The qualitative analysis follows an MQM-inspired annotation framework adapted to the tokenizer and script-level phenomenon. All annotations were performed at the segment level by expert linguists

Table 31: Paired bootstrap comparison against the IndicTrans2 baseline for Hindi→X translation. $\Delta\text{chrF++}$ reports the difference between each model and IndicTrans2, with 95% confidence intervals from paired bootstrap resampling; positive values indicate improvement over the baseline.

Language	Model	$\Delta\text{chrF++}$ [95% CI]	p
Gondi	NLLB	-14.05 [-14.96, -13.23]	<0.001
Gondi	ByT5	+13.67 [+12.62, +14.62]	<0.001
Kui	NLLB	-4.77 [-5.09, -4.46]	<0.001
Kui	ByT5	+9.97 [+9.31, +10.70]	<0.001
Garó	NLLB	-1.80 [-2.71, -0.92]	<0.001
Garó	ByT5	+4.66 [+3.76, +5.55]	<0.001
Mundari	NLLB	+14.60 [+13.73, +15.56]	<0.001
Mundari	ByT5	+10.46 [+9.50, +11.59]	<0.001
Bhili	NLLB	-24.44 [-25.39, -23.54]	<0.001
Bhili	ByT5	+2.28 [+1.19, +3.33]	<0.001
Santali	NLLB	+7.87 [+7.03, +8.76]	<0.001
Santali	ByT5	-15.48 [-16.67, -14.30]	<0.001
Kokborok	NLLB	+16.40 [+15.39, +17.38]	<0.001
Kokborok	ByT5	+13.41 [+12.57, +14.20]	<0.001

Table 32: Paired bootstrap comparison against the IndicTrans2 baseline for X→Hindi translation. $\Delta\text{chrF++}$ reports the difference between each model and IndicTrans2, with 95% confidence intervals from paired bootstrap resampling; positive values indicate improvement over the baseline.

Language	Model	$\Delta\text{chrF++}$ [95% CI]	p
Gondi	NLLB	-22.08 [-23.24, -21.00]	<0.001
Gondi	ByT5	+11.37 [+10.31, +12.44]	<0.001
Kui	NLLB	-12.45 [-13.11, -11.89]	<0.001
Kui	ByT5	+2.75 [+2.02, +3.41]	<0.001
Garó	NLLB	-21.64 [-22.67, -20.67]	<0.001
Garó	ByT5	+0.06 [-0.99, +1.00]	0.896
Mundari	NLLB	+20.56 [+19.57, +21.65]	<0.001
Mundari	ByT5	+14.39 [+13.11, +15.72]	<0.001
Bhili	NLLB	-36.08 [-37.29, -34.90]	<0.001
Bhili	ByT5	-1.93 [-3.13, -0.63]	0.004
Santali	NLLB	+4.16 [+3.30, +5.00]	<0.001
Santali	ByT5	-21.16 [-22.34, -20.07]	<0.001
Kokborok	NLLB	+10.84 [+10.06, +11.56]	<0.001
Kokborok	ByT5	-3.97 [-4.72, -3.27]	<0.001

who were not involved in the original translation process. For each example, annotators aligned the system output with both the source sentence and the human-verified reference translation in order to identify semantic, linguistic, and script-related failures.

Each detected error was assigned (i) a category from the taxonomy in Appendix M, (ii) a sever-

Table 33: Paired bootstrap comparison against the IndicTrans2 baseline for English→X translation. $\Delta\text{chrF++}$ reports the difference between each model and IndicTrans2, with 95% confidence intervals from paired bootstrap resampling; positive values indicate improvement over the baseline.

Language	Model	$\Delta\text{chrF++}$ [95% CI]	p
Gondi	NLLB	-11.46 [-12.17, -10.85]	<0.001
Gondi	ByT5	+3.73 [+2.74, +4.69]	<0.001
Kui	NLLB	+2.05 [+1.60, +2.50]	<0.001
Kui	ByT5	+10.43 [+9.80, +11.03]	<0.001
Garó	NLLB	+17.55 [+16.57, +18.64]	<0.001
Garó	ByT5	+19.12 [+18.18, +20.09]	<0.001
Mundari	NLLB	+12.92 [+11.99, +13.88]	<0.001
Mundari	ByT5	+4.94 [+3.91, +6.02]	<0.001
Bhili	NLLB	-16.61 [-17.36, -16.00]	<0.001
Bhili	ByT5	-2.28 [-3.24, -1.30]	<0.001
Santali	NLLB	+13.33 [+12.10, +14.66]	<0.001
Santali	ByT5	-12.82 [-13.63, -11.98]	<0.001
Kokborok	NLLB	+21.01 [+19.93, +22.10]	<0.001
Kokborok	ByT5	+18.28 [+17.38, +19.13]	<0.001

Table 34: Paired bootstrap comparison against the IndicTrans2 baseline for X→English translation. $\Delta\text{chrF++}$ reports the difference between each model and IndicTrans2, with 95% confidence intervals from paired bootstrap resampling; positive values indicate improvement over the baseline.

Language	Model	$\Delta\text{chrF++}$ [95% CI]	p
Gondi	NLLB	-7.97 [-8.90, -7.03]	<0.001
Gondi	ByT5	+1.66 [+0.62, +2.60]	0.002
Kui	NLLB	-2.50 [-3.14, -1.74]	<0.001
Kui	ByT5	+3.71 [+2.90, +4.53]	<0.001
Garó	NLLB	+5.89 [+5.21, +6.64]	<0.001
Garó	ByT5	+2.58 [+1.84, +3.32]	<0.001
Mundari	NLLB	+9.40 [+8.48, +10.46]	<0.001
Mundari	ByT5	-0.73 [-2.00, +0.61]	0.298
Bhili	NLLB	-8.42 [-9.53, -7.44]	<0.001
Bhili	ByT5	-2.55 [-3.48, -1.59]	<0.001
Santali	NLLB	+2.08 [+1.29, +2.93]	<0.001
Santali	ByT5	-18.14 [-19.02, -17.26]	<0.001
Kokborok	NLLB	+9.43 [+8.75, +10.17]	<0.001
Kokborok	ByT5	+1.03 [+0.39, +1.70]	<0.001

ity score, and (iii) an optional free-text explanation when the predefined category labels did not sufficiently capture the linguistic or cultural nuance of the error. This design allowed the analysis to capture not only surface translation quality, but also failures related to script choice, tokenizer behavior, cultural representation, and language drift that are often obscured by automatic metrics.

Severity scale. Errors were annotated on a five-level severity scale. *Very low* errors correspond to cosmetic issues such as spacing, punctuation, or minor stylistic awkwardness in an otherwise usable translation. *Low* severity marks isolated surface-level problems that are noticeable but do not substantially affect interpretation. *Medium* severity indicates errors that disrupt fluency or content over part of the sentence while leaving the overall meaning recoverable. *High* severity denotes major mistranslations or omissions that render substantial portions of the output unreliable. Finally, *Very high* severity is reserved for unusable outputs, including translations produced in the wrong language, semantically unrelated generations, or character-level corruption.

For aggregate analysis, the five-point scale was further collapsed into three broader bands. *Critical* errors correspond to very high severity and include cases such as complete wrong-language output (E5) and severe tokenization corruption (E8). *Major* errors combine high and medium severity cases, while *Minor* errors group together low and very low severity issues.

Annotation procedure. Annotators recorded every identifiable error within a segment, up to a maximum of five errors per example. When additional issues were present, only the five most consequential errors were retained to maintain consistency across annotations. Error spans were first highlighted and colour-coded before assigning category and severity labels. Errors were annotated at the finest linguistically meaningful granularity. For instance, two independent mistranslations occurring within the same sentence were recorded as separate errors rather than merged into a single label. Where multiple categories overlapped on the same span, only the most severe category was retained. Ties were resolved according to the priority ordering:

Accuracy > Fluency > Terminology > Style.

Source-side problems, including mistranscriptions or references mistakes were marked explicitly on the source segment and excluded from downstream scoring. Outputs that were too corrupted for span-level annotation were labeled as a single *Non-translation* error covering the entire segment.

Two-linguist consensus. Each system output was independently reviewed by two expert linguists with proficiency in the target language. Annotations were accepted directly when both annotators agreed on the severity band and assigned overlapping category labels. Disagreements were resolved through adjudication by a senior language expert, who examined the contested segment and produced the final annotation. This consensus-based procedure was particularly important for endangered and low-resource languages where distinctions between dialectal variation, borrowing, transliteration, and genuine translation errors may be linguistically subtle.

M Human Evaluation Error Taxonomy

The annotation taxonomy comprises eight error categories, each designed to capture a distinct underlying failure mechanism rather than only surface-level errors. The taxonomy combines conventional MT evaluation dimensions such as semantic adequacy and fluency with categories specifically motivated by multilingual tokenization and script representation. In particular, E5 (*Target Language Drift*) and E8 (*Tokenization and Script Artifact*) directly operationalise the central tokenizer-mediated phenomena studied in the paper. The remaining categories capture errors related to semantic transfer, cultural representation, named-entity preservation, source copying, morphology, undertranslation, and robustness. Together, the taxonomy enables a linguistically grounded analysis of how multilingual MT systems fail under low-resource and script-diverse conditions.

E1 Semantic Adequacy. Errors in which the generated meaning diverges from the source sentence. This category includes complete meaning loss, partial omission of key content, semantic distortion, and hallucinated information not present in the source. These failures directly affect translation faithfulness and are often the most severe errors from the perspective of usability.

E2 Cultural Lexical Substitution. Culturally grounded vocabulary replaced with generic or culturally flattened alternatives. Subtypes include kinship terminology, ritual and religious expressions, ecological and agricultural vocabulary, and governance or institutional terms. Although such substitutions may appear acceptable under automatic

metrics, they frequently erase culturally specific distinctions that remain important for community-centered translation settings.

E3 Named Entity Error. Failures involving the preservation or transliteration of named entities, including person names, place names, organizations, institutions, dates, and numerical expressions. This category captures both semantic corruption and script-level instability in entity rendering.

E4 Source-Side Copying. Cases where portions of the source sentence are reproduced rather than translated. Subtypes include complete sentence copying, partial copying of key phrases, and script mixing in which untranslated source-script tokens remain embedded in the output. Such errors are particularly common in closely related language pairs with high lexical overlap.

E5 Target-Language Drift. Outputs generated in a language other than the requested target language. This category includes complete wrong-language generation, substantial unintended code-mixing, and register drift in which Hindi or English lexical patterns dominate endangered-language outputs. Bhili→Gujarati drift under IndicTrans2 represents a paradigmatic example. Within the context, E5 serves as one of the clearest diagnostic signals of vocabulary-proximity bias in multilingual tokenizers.

E6 Morphology and Grammar. Violations of target-language morphology or syntax, including tense errors, agreement failures, incorrect case marking, postposition errors, and word-order distortions. For morphologically rich and agglutinative languages such as Mundari, this category additionally captures the collapse of complex negation–agreement–aspect structures into Hindi-like Devanagari token sequences that obscure native grammatical structure.

E7 Undertranslation and Content Omission. Failures in which information present in the source is omitted or compressed in the output. Subtypes include clause deletion, dropped modifiers, premature truncation, and over-generalized paraphrases that replace specific content with vague formulations. These errors frequently emerge in long or morphologically dense sentences where tokenization length expands substantially.

E8 Tokenization and Script Artifacts. Errors attributable to tokenizer fragmentation, script encoding instability, or vocabulary exhaustion. This category includes character-level corruption, repetitive token loops, malformed Unicode sequences, and mixed-script artifacts such as unexpected Arabic glyphs or Latin diacritics appearing within the output. The English→Kui examples under IndicTrans2 provide concentrated instances of this failure type. Among all categories, E8 most directly reflects breakdowns arising from script-tokenizer interaction and therefore serves as a central diagnostic category for the mechanistic analysis.

N Error Distribution Across the Experimental Grid

We manually analysed a primary sample of 259 IndicTrans2 outputs in the corpus-native script setting. The sample was stratified by language and translation direction, with additional Bhili and Kui outputs included because preliminary audits showed recurring language-drift and script-artifact failures. Comparative samples from NLLB-600M, ByT5-small, prompted LLMs, and the unified-script conditions were inspected separately where relevant.

Overall, 30.5% of outputs were error-free, 20.5% contained only minor issues, 34.7% contained major errors, and 14.3% were critical. Error severity varied substantially across languages. Santali in Ol Chiki was the strongest case, with nearly all sampled outputs judged error-free and only minor undertranslation observed in the remaining cases. Bhili, Mundari, and Gondi showed moderate performance but different failure profiles: Bhili was most affected by language drift and culturally sensitive lexical substitutions, whereas Mundari and Gondi showed more general semantic degradation. Garo and Kokborok had lower error-free rates and frequent semantic adequacy errors, consistent with weaker support for these Latin-script Tibeto-Burman languages in the evaluated subword models. Kui was the most fragile case: no sampled output was fully error-free, and critical failures were concentrated mainly in the En→Kui direction.

Direction also mattered. X→Hindi and X→English were substantially more stable than En→X, while Hi→X remained intermediate. The En→X collapse was driven primarily by generation into weakly represented target languages, es-

pecially Kui, where the decoder often failed to retrieve reliable target-language forms. This pattern supports the tokenizer-mediated account: encoding endangered-language input is difficult but often recoverable, whereas generating endangered-language output exposes sparse or poorly trained regions of the target vocabulary.

Across categories, the dominant errors were semantic adequacy failures, target-language drift, undertranslation, tokenization and script artifacts, named-entity errors, cultural-term substitutions, source copying, and morphological or grammatical loss. These categories were not uniformly distributed: Bhili concentrated drift and cultural-term errors, Kui dominated tokenization and script-artifact failures, and Garo/Kokborok concentrated semantic adequacy errors. Model comparisons further support this diagnosis: NLLB-600M avoided the catastrophic garbling observed for IndicTrans2 on En→Kui but produced sparse Odia-influenced output; ByT5 reduced garbling but undertranslated more often; prompted LLMs failed mainly through drift and source copying.

Script-condition audits were consistent with the automatic results. Bhili Latin-unified outputs under IndicTrans2 showed more semantic and drift errors, matching the large chrF++ decline in Table 3. In contrast, Kui under NLLB-600M shifted from critical garbling toward more recoverable major errors under Devanagari unification, matching the observed chrF++ gain. These patterns indicate that script effects arise from tokenizer behaviour and pretraining coverage rather than fixed properties of the languages themselves.

O Representative Translation Errors

The examples below are drawn from the manually annotated evaluation set and span all major language families represented in the corpus. Each example illustrates a dominant error category from the taxonomy (E1–E8; Appendix M) and connects the observed failure to the broader quantitative patterns reported in the paper. Figures 8–11 reproduce the original system outputs.

O.1 Errors in Devanagari-Native Languages: Gondi, Mundari, and Bhili

Gondi: Critical Domain-Adjacent Hallucination (E1). The source sentence contains *Aal-siram*, a folk compound roughly meaning “the one who never works,” together with a negated action

verb. IndicTrans2 replaces this with *Saasram*, a phonologically similar name sharing overlapping Devanagari subword segments, while simultaneously reversing the negation. The resulting output is semantically unrelated to the source. The same substitution appears in the Hindi→Gondi direction, suggesting a systematic subword-level confusion rather than an isolated decoding error. NLLB, by contrast, produces heavily degraded vowel-stripped and diacriticless output in both directions, consistent with its −22.1 chrF++ deficit relative to IT2 on X→Hindi (Table 32).

Mundari: Morphological Flattening and Lexical Confusion (E6/E1). IndicTrans2 translates “Can you speak?” without catastrophic semantic loss but shifts to the formal pronoun and reframes the utterance as “what do you say?” This produces a register and intent mismatch that weakens the distinction between ability and willingness encoded in the Mundari source. In the reverse direction, NLLB substitutes *jagar* (“quarrel” or “argument”) for *kaji* (“speak”), two orthographically similar forms under Devanagari segmentation. The result introduces an unintended confrontational meaning and alters the communicative force of the sentence.

Bhili: Meaning Distortion and Cultural-Term Substitution (E1/E2). The source expresses devotional gratitude directed toward God. IndicTrans2 reverses the underlying speech act, producing a blessing directed *at* God rather than gratitude expressed *about* God. At the same time, a culturally grounded gratitude expression is flattened into a generic blessing formula. Although the output receives a relatively high automatic score, the example illustrates how lexical overlap can obscure representational loss. IndicTrans2’s strong Bhili-Hindi scores in both directions are partly driven by surface overlap ($Ov=0.608$) rather than by robust semantic transfer (Appendix F). NLLB again produces collapsed, vowel-stripped output, mirroring its large quantitative deficit on Bhili→Hindi.

O.2 Errors in Latin-Native Languages: Garo and Kokborok

Garo: Major Undertranslation under Script Unification (E7). In the native-script setting, IndicTrans2 produces near-reference Garo→English translations for biographical content: named entities are preserved, and word order is correctly re-ordered into English SVO structure despite the ab-

sence of explicit Garo pretraining coverage. Under Devanagari unification, however, the same model drops the culturally informative phrase referring to “intricate stone carvings” while preserving the architectural term *Shikhara*. The degradation is selective rather than global: culturally specific content lacking strong Devanagari subword support is preferentially lost during decoding.

Kokborok: Entity-Chain Truncation and Undertranslation (E7). IndicTrans2 generates fluent English from Kokborok input but omits historically central references to the Madurai Nayakas and the Carnatic region while preserving mentions of Adil Shahi and the Maratha army. The output therefore receives partial chrF++ credit despite dropping a substantial portion of the factual content. The example illustrates a broader limitation of overlap-based metrics: surface fluency can co-exist with severe informational loss. Under Devanagari unification, geographical entities such as Itarsi, Hoshangabad, and Betul remain comparatively stable, indicating that place-name transfer is more robust than the transfer of culturally specific narrative content.

O.3 Errors in Ol Chiki and Odia Script Settings: Santali and Kui

Santali: Semantic Hallucination and Numeral-Script Variation (E1/E3). For the sentence “shopping will be completed by 20 December,” IndicTrans2 differs from the reference only in numeral formatting, using standard Arabic numerals instead of the native-script numeral form, while NLLB remains semantically faithful. ByT5, however, generates fluent text in the correct script but hallucinates an unrelated temporal statement referring to an entire century rather than the intended shopping event. The Santali→English direction reveals the same pattern: IndicTrans2 and NLLB preserve the original meaning, whereas ByT5 produces semantically unrelated content centred on “the 20th century.” Importantly, the failure is not orthographic. ByT5 generates well-formed script sequences, but lacks the subword-level semantic grounding that IndicTrans2 acquires through explicit native-script coverage during pretraining.

Kui: Vocabulary Exhaustion, Repetition, and Cultural-Term Loss (E8/E1/E2). En→Kui under IndicTrans2 fails catastrophically across all tested domains (Figure 11). Outputs contain mixtures of unrelated Unicode symbols, isolated script

diacritics, and fragmented character sequences with no recoverable Kui lexical content. The pattern remains consistent across education, health, agriculture, culture, and public-service examples, indicating a structural vocabulary failure rather than a domain-specific weakness. IndicTrans2 lacks sufficiently dense Kui subword coverage in the Odia script, causing the decoder to collapse into character-level noise during generation.

Kui→Hindi outputs show a clear severity gradient. Public-service sentences remain partially interpretable despite agent loss and paraphrastic simplification, whereas culturally grounded examples collapse into repetition loops. In one folk-tradition example, references to songs, dances, and oral storytelling practices are reduced to repetitive generic phrases, erasing nearly all community-specific content. The single near-reference exception appears under Devanagari unification, where IndicTrans2 reproduces the reference almost exactly. This aligns with the STP=0.396 fragmentation reduction reported in Table 10 and makes the central argument concrete: the dominant failure arises not from linguistic complexity, but from the tokenizer’s inability to represent the language effectively.

P Cultural Representation and Dataset Governance

A central challenge in endangered-language NLP is that linguistic coverage alone does not ensure meaningful representation. Translation systems must preserve not only propositional meaning, but also the cultural, ecological, and social knowledge encoded in community vocabulary, kinship systems, oral traditions, and local governance practices. Our corpus was therefore designed as both a parallel MT benchmark and a documentation-oriented resource that preserves links between language form and community-specific meaning.

The dataset covers culturally grounded domains, including agriculture, ritual practice, oral storytelling, public services, local ecology, and community institutions. The qualitative analysis in Section 6 shows that errors in these domains are systematic rather than incidental: they reflect representational biases introduced by multilingual tokenization and uneven pretraining coverage.

The error classes most directly tied to cultural representation are target-language drift (E5) and cultural-term substitution (E2). Both oc-

cur when limited target-language vocabulary support pushes the model toward a higher-resource neighbour with denser tokenizer coverage. In Bhili, this appears as drift toward Gujarati; in Gondi and Mundari, community-specific kinship and agricultural terms are normalised into Hindi-like paraphrases; and in Garo and Kokborok, fluent outputs often replace culturally grounded expressions with generic English-style approximations. Across these cases, the decoder substitutes broader high-resource expressions for lexical units that are weakly represented in the tokenizer and pretraining data.

These patterns show that vocabulary allocation is not socially neutral. In endangered-language settings, culturally specific substitutions are not merely stylistic differences: they flatten distinctions central to community identity, oral knowledge, and language documentation. Repeated replacement of local terms with high-resource paraphrases risks erasing precisely the variation that revitalisation initiatives seek to maintain.

The corpus design reflects these concerns in three ways. First, native speakers produced the references and expert linguists independently verified them to preserve culturally grounded vocabulary. Second, the annotation protocol treats cultural preservation as an evaluation dimension distinct from general semantic adequacy, allowing culturally meaningful substitutions to be analysed even when automatic scores remain high. Third, the Santali setting highlights the importance of script integrity: script conversion may improve token overlap while weakening orthographic identity and community validity.

Taken together, these choices position the corpus as both an MT benchmark and a computational language-documentation resource. They also reinforce the claim: tokenizer design is not only a technical optimisation problem, but a representational decision that shapes whether multilingual systems can support endangered-language communities faithfully.

Q Generalizing Beyond the Seven-Language Benchmark

The numerical outcomes are conditioned on the seven languages, three model families, and validated script-conversion pathways in our pipeline. We do not assume that the same effect sizes will transfer directly to other language model pairs,

since typology, orthography, conversion tools, and tokenizer design can all change the direction and magnitude of script effects. The transferable contribution is methodological: measure Script Tokenization Parity before fine-tuning, validate non-trivial conversion with speakers where possible, and inspect error types that automatic metrics may understate. The benchmark spans four families and four corpus scripts, providing a controlled setting in which tokenizer-mediated reversal appears across typologically distinct languages. Santali further illustrates the safeguard: even when unification could reduce tokenisation cost, it should not be used without a lossless and community-acceptable conversion pathway.

Q.1 Cross-Regional Case Study: Amharic

To examine whether our findings extend beyond the Indian subcontinent, we conduct a cross-regional case study on Amharic, a Semitic language written in the Ge’ez script in Ethiopia. We use a public ~ 20 k-pair training corpus and a 1 k-sentence test set, following the same fine-tuning and evaluation setup used in our main experiments (Table 35). NLLB-600M, for which Amharic is a supported language, produces more compact Ge’ez segmentation and higher translation scores, whereas ByT5-small exhibits the expected byte-level pattern of substantially longer sequences without relying on a learned subword vocabulary. We treat these tokenization statistics descriptively rather than as a single-factor explanation of translation quality. IndicTrans2 is not evaluated because Amharic and Ge’ez fall outside its intended language and script scope.

We evaluate Amharic only in its native Ge’ez script. Without community validation for a linguistically and culturally acceptable conversion, we do not introduce an alternative-script condition, following the same validity criterion applied to Santali (Section 3). This case therefore provides a cross-regional check, beyond the Indian subcontinent, on the portability of our training and tokenizer-diagnostic workflow, rather than a full external replication of the multi-script ScSens analysis. More broadly, the transferable contribution is methodological rather than an expectation of identical effect sizes: measure STP before fine-tuning, validate non-trivial script conversions with speakers, and inspect errors that surface metrics may understate.

Table 35: Amharic results in the native Ge’ez script. Amharic-side tokenization statistics are reported for Amh→En. Vocabulary utilisation is tokenizer-specific and is not directly comparable between NLLB-600M’s learned subword vocabulary and ByT5-small’s fixed byte inventory.

Model	Dir.	BLEU	chrF++	Fert.	3+ units	Vocab util.
NLLB-600M	Amh→En	41.21	61.45	2.08	32.8%	3.1%
NLLB-600M	En→Amh	21.94	44.15	–	–	–
ByT5-small	Amh→En	17.66	38.85	12.22	97.9%	64.5%
ByT5-small	En→Amh	8.73	25.13	–	–	–

R Dataset Scale and Statistical Robustness

The approximately 20k-sentence scale reflects a realistic data regime for endangered-language MT, where fluent literate speakers are limited, on-line orthography is inconsistent, and community-specific terminology is often absent from web text. We assess robustness through statistical significance tests (Appendix K), full baseline results (Tables 31–34), script-conversion effects relative to the corpus-native baseline (Table 3), and quality-controlled corpus construction with attrition reported in Table 6. The resulting benchmark is smaller than high-resource MT datasets, but combines speaker-verified references, controlled script conditions, and statistical testing in a resource regime realistic for endangered languages.

Translation from Mundari to Hindi : Source : चाना: आम काजि दाड़िआम? Reference : क्या तुम बोल सकते हो? IndicTrans2 output: चाय: आप क्या कहते हैं? NLLB output: क्या आप बात कर सकते हैं? Translation from Hindi to Mundari : Source : क्या तुम बोल सकते हो? Reference : चाना: आम काजि दाड़िआम? IndicTrans2 output: चिनअ: अम कजिओ:अ:? NLLB output: चिनअ: अम जगर दड़िअ:? Translation from Bhili to Hindi : Source : भगवाने भलुं करयूं एम हूँ मानु । Reference : भगवान का सुक्रगुजार हूँ IndicTrans2 output: भगवान अच्छा करें, मैं मानता हूँ। NLLB output: भगवन न भल कर, म मनत ह Translation from Hindi to Bhili : Source : भगवान का सुक्रगुजार हूँ Reference : भगवाने भलुं करयूं एम हूँ मानु । IndicTrans2 output: भगवान नो शुक्रगुजार से। NLLB output: भगवन न सकर मळह Translation from Gondi to Hindi : Source : आलसीराम कुद आस बस्केय बदे बूतो तुन के वोर। Reference : आलसीराम खुद से कभी ही किसी काम को नहीं करता IndicTrans2 output: सासराम ने कहा कि हम शायद ही किसी अन्य कार्य के लिए जा रहे हैं। NLLB output: आलसीराम कड आस वशष क लए ससकर करन चहए Translation from Hindi to Gondi : Source : आलसीराम खुद से कभी ही किसी काम को नहीं करता Reference : आलसीराम कुद आस बस्केय बदे बूतो तुन के वोर। IndicTrans2 output: सासराम आपुना ता कुबैय कुबैय कुबे काज कियाना। NLLB output: आलसीराम लहक नकन बसकैय कैययतरै
--

Figure 8: Devanagari-native languages (Gondi, Mundari, Bhili). IT2 hallucinates a name (*Saasram* for *Aalsiram*) and inverts negation in Gondi→Hindi; NLLB produces vowel-stripped, diacriticless output across all three pairs. Mundari↔Hindi shows recoverable register and aspect errors rather than catastrophic failure, consistent with its intermediate chrF++ scores.

Translation from English to Garo (Native Script) : Source: What is an Equatorial Telescope Mount? Reference : Equatorial Telescope Mount maia? IndicTrans2: Equatorial Telescope Mout minga? Translation from Garo to English (Native Script) : Source : Saund British India ni Punjab Province o Amritsar district ni Chhajulwadi songo saksaprak Indian Punjabi Sikh nokdang o atchiaha. Reference : Saund was born in Chhajulwadi, Amritsar district, Punjab Province, British India, to an Indian Punjabi Sikh family. IndicTrans2 output: Saund was born in an Indian Punjabi Sikh family in Chhajulwadi village of Amritsar district in the Punjab province of British India. Translation from English to Garo (Devanagari Script) : Source : The temple, located near the Police Post and the Treasury building, is built in the Shikhara style, with intricate stone carvings. Reference : पोतिये पोस्त अरो वेअसुय बुइल्लिङ्ग सेपन्गो दोन्निङ्गप इअ मन्दिरको सिंहखर बेअलिख रिक्ह जनों रो ओन्ग रिगो नितोए नोक्सरन्ग सलह. IndicTrans2 output: इअ मन्दिर-नि पुइस पोस्त अरो वेछेरि बिल्दिङ्गको बक्-अह, इअर शिखरा स्तैल्-ओ ज · अ, सोन्ग · अ. Translation from Garo to English (Devanagari Script) : Source : मुल्लि स्प्स्तेम् ओपेर्तोर्-रन्ग मलयलम्, एन्ग्लिस्ह अरो गिपिन् इन्दिअ-नि कुऽसिक्कन्ग बक्स अऽगिल्सिन्ग-नि छन्नेल्-रन्गको ओन्ऽअ. Reference : Multi system operators provide a mix of Malayalam, English, other Indian language and international channels. IndicTrans2 output: Multi-system operators operate channels in Malayalam, English, and other Indian languages. Translation from English to Kokborok (Native Script) : Source : The temple, located near the Police Post and the Treasury building, is built in the Shikhara style, with intricate stone carvings. Reference : O nokthar, police post tei rangchakni nok ganao tongnai, shikhara style bai swnamjak, holongni swnamjak manwirok bai. IndicTrans2 output: Kwrwngmungrokn ni nwio, khorokni tongmani thanikha, abo Shikhara style-o swnamjak, bisingo kwbangma bwthamni mwthairok. Translation from Kokborok to English (Native Script) : Source : Fort complex o Bijapur ni Adil Shahi ni habugra tei Maratha ni habugra ni sengkrak rok ni bisingo belai kheno choba wngwi tongkha. Reference : The fort complex has witnessed fierce battles between the Madurai Nayakas and Adil Shahi dynasty of Bijapur, Carnatic region and Maratha Imperial forces. IndicTrans2 output: The fort complex witnessed a fierce battle between the forces of Adil Shahi of Bijapur and the Maratha army. Translation from English to Kokborok (Devanagari Script) : Source : It is located in Itarsi of Hoshangabad District of Madhya Pradesh state, above Betul district. Reference : अबो मध्य प्रदेश हस्ते नि होस्हन्गबद् अम्छैनि इतर्सि ओ, बेतुल् अम्छैनि सकओ तोन्गो. IndicTrans2 output: ओ मध्य प्रदेश हक्त् wइनि होस्हन्गबद् दिस्त्रिक् नि इतर्सि ओ, बेतुल् दिस्त्रिक् नि थनि तोन्गो. Translation from Kokborok to English (Devanagari Script) : Source : अबो मध्य प्रदेश हस्ते नि होस्हन्गबद् अम्छैनि इतर्सि ओ, बेतुल् अम्छैनि सकओ तोन्गो. Reference : It is located in Itarsi of Hoshangabad District of Madhya Pradesh state, above Betul district. IndicTrans2 output: It is located in the Itarsi of Hoshangabad district, Madhya Pradesh, near Betul district.
--

Figure 9: Latin-native languages (Garo, Kokborok) under native and Devanagari-unified conditions. Native-script Garo→English (IT2) is near-reference; named entities are preserved and word order is faithfully reordered. Devanagari-unified conditions reveal incomplete cultural-content transfer: stone carvings and international channels are dropped (E7), and entity chains are truncated in Kokborok→English (Madurai Nayakas, Carnatic region absent from output).

Source: For example, shopping will be completed by 20 December.
 Reference : ඉනි.ලබන. පව්වන02, 20 තැවැදුමයක් යාමයක් වනාහි යාමය.ඉනි.ලබන.පව්වන02
 IndicTrans2 output : ඉනි.ලබන. පව්වන02, 20 තැවැදුමයක් යාමයක් වනාහි යාමය.ඉනි.ලබන.පව්වන02
 ByT5 output : ඉනි.ලබන.පව්වන02, 20 තැවැදුමයක් යාමයක් වනාහි යාමය.ඉනි.ලබන.පව්වන02
 NLLB output : ඉනි.ලබන. පව්වන02, 20 තැවැදුමයක් යාමයක් වනාහි යාමය.ඉනි.ලබන.පව්වන02

Source: ፪፻፳፱ ዓ.ም. ፳፱ ወርሀ ሐምሌ ፲፱፻፲፱ ዓ.ም. ስራ ይጠናቃል
Reference: For example, shopping will be completed by 20 December.
IndicTrans2 output: For example, the purchase will be finished by December 20.
ByT5 output: According to the state, the 20th century work is the state of the 20th century.
NLLB output: According to the report, the sale will be completed by December 20.

Source : अपने घर - परिवार का ध्यान रखा करो ।
Reference : ନିଜା ଇତ୍ତୁ – କୁମବୁଝିତାରା ଏଲ୍ଲା ଭାଙ୍ଗାମ୍ ।
IndicTrans2 output: ବେସାରି ମାକା କନା ପାରାନ ସିକା ବାବା ସ୍ୱାମୀ ।
ByT5 output : ନନ୍ଦ ଇତ୍ତ ଉଡ଼କମବ ଭଙ୍ଗମ୍
NLLB output : ଇତ୍ତବୁଝିତା ଇତ୍ତବୁଝିତା ନିଗୁରଇ ମ୍ୟାସାସ୍ତ୍ର

Source : [एश्याङ्ग घोषालय एका जठरा कर्माचार्या आर्वाघेक](#) ।
 Reference : उन्होंने पुलिस और वाइल्ड लाइफ के कर्मचारियों को बुलाया।
 IndicTrans2 output : उन्होंने पुलिस और जय्ये के कर्मचारियों से बात की।
 ByT5 output : वह पोलिस और जाति कर्मचारी के लिए पोलिस और जाति कर्मचारी के लिए है।
 NLLB output : उनहन यह बतय क वह एक जनजत के नयुय थ

Source : sarojīni nagara mārkēṭa tṛedārsa āsōsiesanata sabhāpātī aśoka randhāvā besāneṛjū āmu belene ṛahanā deheṅgi dukānī jāpīnamū .
 Reference : सरोजिनी नगर मार्केट ट्रेडर्स एसोसिएशन के अध्यक्ष अशोक रंधावा ने कहा कि , '1 हम पहले की ही तरह दुकानें खोल रहे हैं .
 IndicTrans2 output: सरोजिनी नगर मार्केट ट्रेडर्स एसोसिएशन के अध्यक्ष अशोक रंधावा ने कहा कि हम एक साथ काम कर रहे हैं।
 ByT5 output : सरोजिनी की नगर मार्केट ट्रेडर्स एसोसिएशन ने कहा कि हम पहले से पहले से संबंधित करने के लिए सभी के ब
 NLLB output : सरोजिनी नगर बाजार ट्रेडर्स एसोसिएशन के सहपाठी अशोक रंधावा ने कहा कि हम इस अभियान को आगे बढ़ाते हैं।

Source : एआज्जु ताडा मोबाइल फोन्डाइ पोलिसइ बेङ्ग्रासेज्जु ।
 Reference : उसने अपने मोबाइल फोन से पुलिस को फोन किया।
 IndicTrans2 output : उसने अपने मोबाइल फोन से पुलिस को फोन किया।
 ByT5 output : उन्होंने अपने मोबाइल फोन से पुलिस को बताया।
 NLLB output : उसने अपना मोबाइल फोन पुलिस को दे दिया।

39147

Source : We have books to read from school.
Reference : ଆମ ଗରବ ଲଡ଼େଇ ଆକ ଡମବ ସଲ ପଡ଼େଗ ପଣସଦନମ
Model output: ଓଁ ଣ଼ ଡ଼଼® Nob'âତିଆ'âT'î'ବା'Â' â'କଷଆ

Source : It is engaged in the development of Ayurvedic medicines for the treatment of animal diseases.
Reference: ଓଡ଼ିଶା ଲୋକ ଶାସ୍ତ୍ର ଲେଖକଙ୍କ ଦ୍ଵାରା ପ୍ରାଚୀନ ଶାସ୍ତ୍ର ଓଡ଼ିଶା ଶାସ୍ତ୍ରମାନେ ।
Model output: ଓଡ଼ିଶା ଲୋକ ଶାସ୍ତ୍ର ଲେଖକଙ୍କ ଦ୍ଵାରା ପ୍ରାଚୀନ ଶାସ୍ତ୍ର ଓଡ଼ିଶା ଶାସ୍ତ୍ରମାନେ ।

Source : ଇ ଗାମପତରା ପୁଣ୍ଡୁ ବାନ୍ଦବାରା ଗୁଲେ ପୁଣ୍ଡୁଡ଼ିନି ଅଧୀନାଲ ଇଆଡେଂଡ଼ିଟି ଟଣ୍ଡା ମାନେଲୁ । ଏମ୍ବେରିତାରି ପାଦା ଆଇନେ 'ଅପରେସନ୍ ମାଲ୍ୟୁଧ' ।
Reference: This group of officers along with their entire team has started a task called 'Operation Malyudh' .
Model output: He also said that the government has decided to launch an **operation** in this regard.

Source : ମୁଣ୍ଡା ଲ'କୁଡାରି ସା'ବା ବାବୁବା ବାବୁବିଜା ଗା'ମେ ଗା'ଡ଼ିଙ୍ଗା ଏଙ୍ଗୁଆ ଏ'ନ୍ଦା ଆତାନା ମାନେ ।
Reference : The traditional festivals of the Munda people feature **special types of music and dance**.
Model output: The Munda people are also known for their **religious beliefs and beliefs**.

Source : मुण्डा लोग के पारंपरिक त्योहारों में विशेष प्रकार के संगीत और नृत्य का आयोजन होता है।
Reference : पूर्वा ळ्कुता वि ान्ना वल्का वल्किता ण्णे ढाडिङ्गा एङ्गा एङ्गा यागाना मने ।
Model output: पू'ना वा'ना वा'या'ना कवि वेङ्गा ळ्क विषमनि घ'मा'ना, मा'ला वा'या वि' ये'य'वा' नावने, वा'हि ये'य' विने

Source : ଝରତା ଜନଜାତି କେ ସେଇସ୍ୟ ପରିବାର କୌ କୃଷି, ସ୍ବାସ୍ଥ୍ୟ, ଭୂମି, ପଶୁ ବ୍ୟବସାୟ ଉତ୍ପାଦ ଆଦି କେ ଲିପି ଶ୍ରୁଣ କମାତେ ହିଁ ।
Reference : ବରଜା ଲଗତ ଲକ ପରବର, ତସ, ସାପସାସ, କତଡ଼ା କତଡ଼ା କତକନ ଲଗ ପନ
Model output: ଲାବାବା ପାକା ସା ନା ମୋଡ଼ାରି ସିଂଗା ବରାଜି ମାରାରି, ବିରାଜା, ସରୀ, ରାକି ପାଂଗା ବାରେ, କଲା କାପି ବେବାବା ରାମି ହାରା

Source : କୋରାଓ (ଡିହାରୀ) ଲଂକୁଲୁଲ ଲଂକୁ ଯା'ନା-ଦୁକାରିତା ଲଂକୁଗାଣି, ଲଂକୁ'ଫ'ନଦାକରୋହା, ଲଂକୁକେରଣି ଇତ୍ୟାଦି ଥାନ୍ତି ।
Reference : कोरवा (डिहारी) जनजाति के लोग परम्पराओं में लोग गीत, लोक नृत्य, लोक कथाएं आदि प्रचलित है ।
Model output: कोरवा (डिहारी) के लोग संस्कृति में लोग हैं, लोग हैं और लोग हैं।

Source : कक्षा विशेषज्ञकृतितु पुस्तक थाका कक्षाविषय गुरागुराकितिता गाठितानिवाहा कनपठिता गुरागुराविवा कनित गुरुतु विवाथाकाना ।
Reference -इस विज्ञान को समझते हुए ही, राष्ट्रीय शिक्षा नीति में, स्थानीय भाषा में, पढाई पर जोर दिया गया है।
Model output: इस कार्यक्रम को पुरा करने के लिए राष्ट्रीय स्तर पर शिक्षा के क्षेत्र में शिक्षा के महत्व पर जोर दिया गया है।

39148

Table 36: Full tokenisation statistics by language, script condition, and tokenizer. Fertility is the mean number of subword tokens per whitespace token, and pieces per sentence is the mean subword count per sentence, equal to fertility times the mean sentence length reported in Table 7. The subword tokenisation parity is computed relative to the corpus-native condition for the same language and tokenizer. Hindi overlap is the proportion of produced subword types that also occur in the Hindi pivot under the same tokenizer.

Language	Script condition	Tokenizer	Fertility	Pcs/sent.	Subword parity	Hindi overlap
Bhili	Original	IndicTrans2	1.3830	24.1	1.0000	0.6084
Bhili	Original	NLLB	1.6893	29.4	1.0000	0.7807
Bhili	Original	ByT5	12.5156	217.6	1.0000	0.9843
Bhili	Devanagari	IndicTrans2	1.3830	24.1	1.0000	0.6084
Bhili	Devanagari	NLLB	1.6893	29.4	1.0000	0.7807
Bhili	Devanagari	ByT5	12.5156	217.6	1.0000	0.9843
Bhili	Latin	IndicTrans2	3.1056	54.0	2.2456	0.0603
Bhili	Latin	NLLB	2.3949	41.6	1.4177	0.0830
Bhili	Latin	ByT5	6.9407	120.7	0.5546	0.6423
Garó	Original	IndicTrans2	3.0453	34.2	1.0000	0.0359
Garó	Original	NLLB	2.5570	28.7	1.0000	0.0364
Garó	Original	ByT5	7.9437	89.3	1.0000	0.7818
Garó	Devanagari	IndicTrans2	3.4722	39.0	1.1402	0.1343
Garó	Devanagari	NLLB	4.1388	46.5	1.6186	0.2127
Garó	Devanagari	ByT5	21.3680	240.2	2.6899	0.5217
Garó	Latin	IndicTrans2	3.0453	34.2	1.0000	0.0359
Garó	Latin	NLLB	2.5570	28.7	1.0000	0.0364
Garó	Latin	ByT5	7.9437	89.3	1.0000	0.7818
Gondi	Original	IndicTrans2	1.8260	17.3	1.0000	0.4368
Gondi	Original	NLLB	2.2185	21.0	1.0000	0.6301
Gondi	Original	ByT5	14.3373	135.8	1.0000	0.8947
Gondi	Devanagari	IndicTrans2	1.8260	17.3	1.0000	0.4368
Gondi	Devanagari	NLLB	2.2185	21.0	1.0000	0.6301
Gondi	Devanagari	ByT5	14.3373	135.8	1.0000	0.8947
Gondi	Latin	IndicTrans2	3.4478	32.7	1.8882	0.0668
Gondi	Latin	NLLB	2.5779	24.4	1.1620	0.0726
Gondi	Latin	ByT5	7.6012	72.0	0.5302	0.6879
Kokborok	Original	IndicTrans2	2.6255	44.9	1.0000	0.0323
Kokborok	Original	NLLB	2.1816	37.3	1.0000	0.0363
Kokborok	Original	ByT5	6.7441	115.4	1.0000	0.6923
Kokborok	Devanagari	IndicTrans2	3.0302	51.8	1.1542	0.1407
Kokborok	Devanagari	NLLB	3.5259	60.3	1.6162	0.2241
Kokborok	Devanagari	ByT5	18.0122	308.2	2.6708	0.5127
Kokborok	Latin	IndicTrans2	2.6255	44.9	1.0000	0.0323
Kokborok	Latin	NLLB	2.1816	37.3	1.0000	0.0363
Kokborok	Latin	ByT5	6.7441	115.4	1.0000	0.6923
Kui	Original	IndicTrans2	6.0472	71.9	1.0000	0.0101
Kui	Original	NLLB	3.1215	37.1	1.0000	0.0438
Kui	Original	ByT5	16.0611	190.8	1.0000	0.7278
Kui	Devanagari	IndicTrans2	2.3949	28.5	0.3960	0.2073
Kui	Devanagari	NLLB	2.9150	34.6	0.9339	0.3187
Kui	Devanagari	ByT5	15.8684	188.5	0.9880	0.7325
Kui	Latin	IndicTrans2	4.0568	48.2	0.6709	0.0508
Kui	Latin	NLLB	3.0682	36.5	0.9829	0.0579
Kui	Latin	ByT5	8.6213	102.4	0.5368	0.7312
Mundari	Original	IndicTrans2	2.0891	31.8	1.0000	0.5735
Mundari	Original	NLLB	2.4727	37.6	1.0000	0.7665
Mundari	Original	ByT5	14.2142	216.3	1.0000	0.9371
Mundari	Devanagari	IndicTrans2	2.0891	31.8	1.0000	0.5735
Mundari	Devanagari	NLLB	2.4727	37.6	1.0000	0.7665
Mundari	Devanagari	ByT5	14.2142	216.3	1.0000	0.9371
Mundari	Latin	IndicTrans2	3.2942	50.1	1.5768	0.1211
Mundari	Latin	NLLB	2.5993	39.6	1.0512	0.1664
Mundari	Latin	ByT5	7.4915	114.0	0.5270	0.7716

continued on next page

Table 36: continued from previous page

Language	Script condition	Tokenizer	Fertility	Pcs/sent.	Subword parity	Hindi overlap
Santali	Original	IndicTrans2	1.2814	26.9	1.0000	0.0236
Santali	Original	NLLB	2.8457	59.7	1.0000	0.0597
Santali	Original	ByT5	14.1513	296.8	1.0000	0.8639

Note. Original denotes the corpus-native script condition. Devanagari and Latin denote audited conversion conditions. Where the conversion target coincides with the corpus-native script, the two rows are identical by construction. Santali is reported only in its native Ol Chiki condition.