Lost in Tokenization: Rethinking Script Invariance in COMET-based Neural Machine Translation Evaluation

Anonymous ACL submission

Abstract

COMET is widely used as the primary quality signal for non-English machine translation, but its implicit *Script Invariance* assumption that adequacy is reflected regardless of which script encodes the target is rarely tested. We treat romanisation as a controlled causal intervention on IndicMT Eval dataset: rewriting target text into Latin script preserves all linguistic content and isolates the effect of orthographic form. Romanisation eliminates 93.3% of between-language COMET variance, confirming script identity, not language family, as the dominant factor behind cross-language differences. The same intervention collapses COMET-human correlation by 29 to 62% and inflates inference cost by up to $5.57\times$. We trace this to an inversion between Tokenization Parity (TP) and Information Parity (IP): romanisation simultaneously inflates tokenisation overhead and reduces semantic density per token. We introduce three inference-time diagnostics: Script Bias Index (SBI), IP Parity Index (IPI), and Computational Tax, which quantify, zone-classify, and cost this failure without retraining. We recommend reporting them alongside COMET for non-Latin-script targets.[1]

1 Introduction

Modern MT evaluation has shifted from surface metrics like BLEU (Papineni et al., 2002) to neural metrics like COMET (Rei et al., 2020), which embed source, hypothesis, and reference into a shared multilingual space and score quality via cosine similarity. Freitag et al. (2021) demonstrated on high-resource European pairs that COMET correlates with human Multidimensional Quality Metrics (MQM; Lommel et al., 2014) judgements far more strongly than surface metrics, establishing MQM as the gold standard against which COMET is evaluated. On the IndicMT Eval (Sai et al., 2023) corpus we observe the same pattern for five low-resource Indic languages[2] (Spearman ρ^+ of 0.422 to 0.651; §4), confirming COMET's dominance as the de facto standard[3].

The Script Invariance assumption. Every cross-lingual COMET comparison rests on an implicit *Script Invariance* (SI) assumption: the metric should reflect translation adequacy regardless of which script encodes the target. Let $\pi_p(\cdot)$ denote a transliteration that re-encodes a string into script p while preserving its linguistic content. SI requires that for any valid target script p,

$$\mathrm{COMET}(s,\, \pi_p(t),\, \pi_p(r)) = \mathrm{COMET}(s,\, t,\, r). \tag{1}$$

Romanisation gives a natural test with $p = \mathrm{Latin}$: re-encoding target text into Latin script preserves all linguistic content, isolating script as the sole varying factor. Our experiments (§3) apply this intervention, and we find that Eq. 1 fails systematically. Despite this, SI is rarely studied as an independent causal variable. COMET's encoder, XLM-RoBERTa (Conneau et al., 2020), was pre-trained on predominantly Latin-script text with a 250,000-token vocabulary built via Byte-Pair Encoding (BPE), which merges frequent character pairs into subword units (Sennrich et al., 2016). Indic scripts are under-represented in this vocabulary and are heavily fragmented by it (Petrov et al., 2023; Kanjirangat et al., 2025). Heavy fragmentation dilutes the pooled sentence embedding, so cosine similarity between hypothesis and reference no longer reliably tracks translation adequacy. Table 1, drawn from IndicMT Eval, makes the consequence stark — (A): the same translation loses 33.8 COMET points when its surface script is changed.

[1] Data and code: https://anonymous.4open.science/r/script-bias-comet-indic-mt-6FEF/.

[2] ISO codes: HIN=Hindi, MAR=Marathi, GUJ=Gujarati, MAL=Malayalam, TAM=Tamil, DEU=German, SPA=Spanish, ENG=English.

[3] ρ^+ marks Spearman correlations with $p < 0.05$ and ρ^- those not statistically significant. All tests are sentence-level on the IndicMT Eval dataset.

(A) Same translation, two scripts

"Police said they suspect an alleged Daesh militant of responsibility for the attack."

Condition	Human	COMET
GUJ native script	18/25	92.8
Same text, romanised	18/25	59.0

(B) Equal quality, different scripts

"Goma is a tourist city of the Democratic Republic of Congo in the extreme east near Rwanda."

Language / Script	Human	COMET
GUJ (Gujarati script)	25/25	95.8
HIN (Devanagari)	25/25	38.2

Table 1: Two instances of orthographic sensitivity from IndicMT Eval. Human scores are MQM annotations: (A) romanisation is a script-only transformation, so the same score applies to both rows; (B) both targets are independently rated error-free, isolating script as the only varying factor.

(B): two perfect translations differ by 57.6 points solely due to orthographic form.

Prior failure-mode studies of COMET concern language mismatch and translationese (Zouhar et al., 2024) or within-script linguistic perturbations (Karpinska et al., 2022); none varies orthographic form as a controlled condition. The IndicMT Eval authors observed cross-language COMET variation but attributed it to language-family effects rather than testing script. Encoder retraining (Wang et al., 2024; Falcão et al., 2024) neither isolates the causal role of script nor provides diagnostics for existing deployments.

Two gaps remain: **(A)** script is rarely studied as a controlled causal variable; **(B)** no inference-time diagnostic framework exists to quantify the distortion without retraining. We address **(A)** with a controlled romanisation intervention on IndicMT Eval (7K MQM-annotated sentences) that holds source text and human annotations fixed and varies only the target script, so any resulting COMET change is attributable exclusively to script identity. We address **(B)** with three inference-time diagnostics derived from prior measures of Tokenization Parity (TP) and Information Parity (IP) (Petrov et al., 2023; Tsvetkov and Kipnis, 2024).

Contributions: (i) the first controlled test of whether script identity distorts COMET scores independently of translation quality; (ii) three novel diagnostics (SBI, IPI, Computational Tax; §5) that adapt prior representational measures to MT-evaluation metric validation, applicable without retraining; (iii) empirical evidence that romanisation degrades COMET–human correlation by 29 to 62% and raises inference cost up to 5.57×, motivating encoder-level remediation over post-hoc normalisation.

2 Background and Related Work

Tokeniser unfairness and parity indices. The Latin-script bias is directly measurable. Petrov et al. (2023) show tokeniser disparities reach 15× across languages; they also define **Tokenization Parity** (TP), the ratio of subword tokens used per word in a target language relative to English. Tsvetkov and Kipnis (2024) define **Information Parity** (IP), the ratio of a language model's compression efficiency on a target language to its efficiency on English, showing IP predicts downstream performance better than fragmentation counts alone. Kanjirangat et al. (2025) extend both TP and IP to dialectal NLP, confirming tokenisation-linked instability across scripts. Two further measures act as sanity checks for our analysis. **Byte Premium** (Arnett and Bergen, 2025) is the bytes-per-word ratio of target to English; it rules out UTF-8 encoding overhead as a cause. **MATTR** (Covington and McFall, 2010) is the count of unique word forms in a sliding 500-token whitespace window; being vocabulary-free, it is independent of XLM-R and rules out training-corpus artefacts. Together they show the bias comes from the tokeniser itself, not from character encoding or training data (Appendix E).

TP and IP each capture only one dimension of the encoder's burden on a target language: TP measures fragmentation, IP measures semantic density. All three diagnostics we introduce in §5 are functions of the *interaction* between TP and IP: **SBI** combines them into a single burden metric, **IPI** converts IP into actionable reliability zones, and **Computational Tax** quantifies the multiplicative overhead they jointly impose under romanisation. We provide empirical evidence that this interaction explains COMET's observed cross-language behaviour, including the Gujarati anomaly where TP alone misranks the most affected language.

3 Experimental Setup

Datasets. IndicMT Eval (Sai et al., 2023) provides 1,400 MQM-annotated sentences per Indic language, labelled across eleven error types and three severity levels, with identical English sources translated into all five targets to allow direct cross-language comparison. To verify our findings extend to higher-resource Latin-script languages, we

Language	COMET				TP			IP		
	NAT	ROM	Δ_{COMET}	$\rho_{\text{nat}} \rightarrow \rho_{\text{rom}}$	NAT	ROM	Δ_{TP}	NAT	ROM	Δ_{IP}
GUJ‡	**85.70**	**70.61**	**−15.09**	$\mathbf{0.590^{+}} \rightarrow \mathbf{0.341^{+}}$	**1.388**	**1.599**	**+15.2%**	**0.438**	**0.289**	**−34.0%**
TAM	84.88	77.72	−7.15	$0.629^{+} \rightarrow 0.240^{+}$	1.319	2.347	**+77.9%**	0.545	0.174	**−68.0%**
MAL	84.08	70.07	−14.01	$0.651^{+} \rightarrow 0.460^{+}$	1.248	1.847	+48.0%	0.549	0.197	−64.1%
MAR†	71.92	72.94	+1.02	$0.487^{+} \rightarrow 0.278^{+}$	1.165	1.640	+40.7%	0.490	0.280	−42.9%
HIN	71.75	68.32	−3.43	$0.422^{+} \rightarrow 0.230^{+}$	1.210	1.702	+40.7%	0.636	0.265	−58.3%

Table 2: Mean COMET, metric–human Spearman ρ, and the TP–IP Inversion. ‡**The Gujarati anomaly**: largest COMET collapse despite smallest TP inflation, explained by GUJ's uniquely low native IP and high SBI (3.45; §5). †MAR is the sole language where mean COMET rises despite every quality indicator worsening. $^{+}p < 0.05$.

add two WMT24 (Kocmi et al., 2024) controls: ENG-SPA (morphologically simple Latin baseline) and ENG-DEU (morphologically complex Latin counterpoint).[4]

Romanisation as a causal intervention. We apply IndicXlit (Madhani et al., 2023), a deterministic neural transliteration system, to target-side IndicMT Eval text. The intervention preserves all phonological, morphological, and semantic content while changing only orthographic form. Source sentences and MQM scores are held fixed across all conditions, so any resulting COMET change is attributable exclusively to script identity. Romanised Indic content is also widely used in everyday online communication (Chavan et al., 2023), so the experiment also doubles as an audit of metric behaviour on a real text genre.

4 The Effect of Romanisation

We first record three empirical observations, then introduce diagnostics that explain why COMET behaves as it does under the intervention.[5] Full per-language statistics are in Appendix A.

Observation 1: Script explains 22.9% of COMET variance. We use one-way ANOVA to quantify how much of the variance in COMET is attributable to script identity treated as a categorical factor. Grouping target languages by script family (Devanagari: HIN, MAR; non-Devanagari: GUJ, MAL, TAM) gives $\eta^2 = 22.9\%$ on native-script COMET. Romanisation reduces this to 1.5% (a 93.3% reduction; Table 2). A natural HIN–GUJ experiment, two closely related Indo-Aryan languages that differ primarily in script, collapses 84% of their native-script gap under romanisation (13.95 → 2.29 points). Script is the dominant factor behind cross-language COMET differences.

[4]Drawn from https://github.com/google-research/mt-metrics-eval

[5]Throughout the paper, subscripts $_{\text{nat}}$ and $_{\text{rom}}$ denote native-script and romanised conditions respectively.

Language	COMET	BLEURT	BERTScore	BLEU	ChrF	TER
GUJ_{nat}	$\mathbf{+0.590^{+}}$	$+0.522^{+}$	$+0.438^{+}$	$+0.339^{+}$	$+0.401^{+}$	-0.327^{+}
GUJ_{rom}	$+0.341^{+}$	$+0.339^{+}$	$+0.375^{+}$	$+0.340^{+}$	$+0.407^{+}$	-0.333^{+}
HIN_{nat}	$\mathbf{+0.422^{+}}$	$+0.381^{+}$	$+0.274^{+}$	$+0.202^{+}$	$+0.224^{+}$	-0.236^{+}
HIN_{rom}	$+0.230^{+}$	$+0.294^{+}$	$+0.214^{+}$	$+0.204^{+}$	$+0.217^{+}$	-0.213^{+}
MAL_{nat}	$\mathbf{+0.651^{+}}$	$+0.596^{+}$	$+0.566^{+}$	$+0.470^{+}$	$+0.542^{+}$	-0.451^{+}
MAL_{rom}	$+0.460^{+}$	$+0.443^{+}$	$+0.529^{+}$	$+0.409^{+}$	$+0.549^{+}$	-0.466^{+}
MAR_{nat}	$\mathbf{+0.487^{+}}$	$+0.443^{+}$	$+0.354^{+}$	$+0.261^{+}$	$+0.292^{+}$	-0.255^{+}
MAR_{rom}	$+0.278^{+}$	$+0.286^{+}$	$+0.313^{+}$	$+0.256^{+}$	$+0.307^{+}$	-0.258^{+}
TAM_{nat}	$\mathbf{+0.629^{+}}$	$+0.563^{+}$	$+0.459^{+}$	$+0.320^{+}$	$-0.312^{\S}$	$+0.410^{\S}$
TAM_{rom}	$+0.240^{+}$	$+0.377^{+}$	$+0.412^{+}$	$+0.286^{+}$	$+0.408^{\S}$	$-0.308^{\S}$

Table 3: Metric–MQM Spearman ρ. Bold: highest $|\rho|$ per language. Neural metrics lose 29–62% under romanisation; surface metrics are flat. TER is negatively oriented. $^{+}p < 0.05$. §TAM ChrF/TER signs flip on the native side from reference n-gram fragmentation.

Observation 2: the TP–IP Inversion. Romanisation *inflates* TP and *collapses* IP in every Indic language (Table 2; TAM at the extreme: TP +77.9%, IP −68.0%). At the language level (n=5), the TP–IP anti-correlation strengthens from $r^{-} = -0.481$ in native script to $r^{+} = -0.896$ under romanisation. The explosion is driven by Tamil digrams (*zh, dh, gh*) and Indic dependent vowels that BPE shatters into near-zero-information tokens (Brahma et al., 2025); the joint shift directly degrades COMET's discriminative power, motivating the three diagnostics in §5.

Observation 3: COMET is uniquely fragile under romanisation. Under native-script conditions, COMET has the highest $|\rho|$ with MQM in every language (ρ^{+} of 0.422 to 0.651; Table 3). After romanisation, COMET drops most steeply in every language ($|\Delta\rho| = 0.19$ to 0.39), exceeding BLEURT (0.09 to 0.19), BERTScore (0.04 to 0.06), and surface metrics (BLEU, chrF, TER; all $|\Delta\rho| < 0.07$; Table 3). The failure is therefore inside the encoder. These observations document the failure but do not give a practitioner anything to compute; the next section operationalises them into three inference-time diagnostics.

5 Diagnostic Metrics

We now introduce the three diagnostics summarised in Table 4; their axiomatic derivations appear in Appendix G. $\overline{\text{TP}}$ and $\overline{\text{IP}}$ denote corpus-level means. TP and IP capture only one dimension each, and the GUJ anomaly is the clearest evidence that they must be combined.

Script Bias Index (SBI): $\text{SBI} = \mathbb{E}_i[\text{TP}_i/\text{IP}_i]$ (where i indexes corpus sentences) is the expected number of subword pieces the encoder spends per unit of information recovered. Its native-script ranking over the five Indic languages reproduces the COMET ranking exactly ($\rho^+ = 1.00$), and it resolves the GUJ anomaly: TP alone places GUJ as the safest of the five, while $\text{SBI}_{\text{nat}} = 3.45$ correctly flags it as the most biased. We propose $\text{SBI} \geq 3.0$ as a practical unreliability flag; the ratio form is justified axiomatically in Appendix G.

IP Parity Index (IPI): $\text{IPI} = |\text{IP} - 1.0|$ converts the continuous IP scale into an actionable distance from English-equivalent parity. Three empirical zones emerge from our seven settings (Figure 1): **Parity** ($\text{IPI} < 0.05$): COMET scores are reliable, with SPA at $\text{IPI} = 0.009$. **Burden** ($0.35 \leq \text{IPI} \leq 0.70$): COMET still tracks adequacy but its absolute scores are depressed by a roughly constant amount; DEU and all Indic native settings sit here. **Paradox** ($\text{IPI} > 0.70$): COMET–human alignment collapses below $\rho = 0.40$. Every Indic language crosses from Burden into Paradox under romanisation, so IPI directly identifies when a COMET score is no longer interpretable as a quality signal. In practice, a deployment that reports COMET on a non-Latin-script target can compute IPI once per language pair: Parity-zone scores can be trusted directly, Burden-zone scores should be reported with IPI attached so absolute values are not over-interpreted, and Paradox-zone scores should be flagged or discarded. The 0.70 boundary is empirically calibrated (Appendix G).

Computational Tax: Romanisation imposes two costs that cannot be traded off: the encoder uses more tokens per word (*Length Penalty*, $\text{LP} = \overline{\text{TP}}_{\text{rom}}/\overline{\text{TP}}_{\text{nat}}$) and each token carries less information (*Entropy Penalty*, $\text{EP} = \overline{\text{IP}}_{\text{nat}}/\overline{\text{IP}}_{\text{rom}}$). Their product $\text{Tax} = \text{LP} \times \text{EP}$ measures how much more encoder compute a romanised sentence demands relative to its native form. Tax ranges from $1.75\times$ (GUJ) to $5.57\times$ (TAM), so a romanised TAM sentence costs over five times its native counterpart to encode while delivering a less reliable score. Since $\ln(\text{Tax}) = \ln(\text{LP}) + \ln(\text{EP})$, EP contributes 62 to 75% of $\ln(\text{Tax})$ across the five Indic languages (Appendix C); per-token information loss, not token-count inflation, is therefore the principal cost, and any remedy must improve the encoder's per-token semantics rather than only reducing fragmentation.

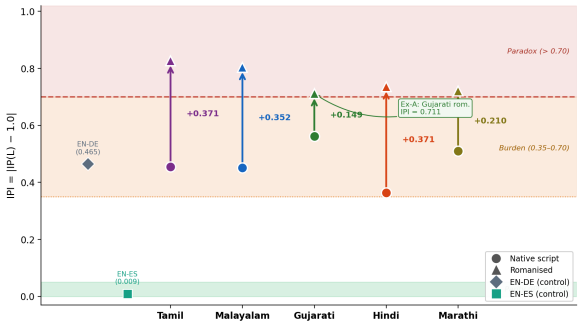

Figure 1: IPI zone transitions under romanisation. Bands: Parity (< 0.05), Burden (0.35–0.70), Paradox (> 0.70). Every Indic language crosses into Paradox after romanisation; ENG-DEU (Burden) and ENG-SPA (Parity) are Latin-script reference points.

Metric	Ideal	Role
$\text{TP} = \lvert t(s_L)\rvert / \lvert t(s_E)\rvert$	1.0	Fragmentation cost; > 1 over-fragmentation.
$\text{IP} = \text{NLL}(w_E)/\text{NLL}(w_L)$	≈ 1.0	Semantic density per token; $\ll 1$ collapse.
$\text{SBI} = \mathbb{E}_i[\text{TP}_i/\text{IP}_i]$	1.0	Burden per unit information. Flag ≥ 3.0.
$\text{IPI} = \lvert\text{IP} - 1.0\rvert$	0.0	Distance from parity; Parity/Burden/Paradox zones.
$\text{Tax} = \text{LP} \times \text{EP}$	1.0	Romanisation overhead; LP, EP defined in §5.

Table 4: Formal definitions of the five diagnostics. $t(\cdot)$ is XLM-RoBERTa's tokeniser; NLL is total negative log-likelihood under BLOOM-560m. **Ideal** denotes the value at English-equivalent parity ($\text{TP} = \text{IP} = 1$).

6 Conclusion

COMET's Script Invariance assumption (Eq. 1) fails systematically: script identity alone explains 22.9% of native-script COMET variance. Cross-lingual COMET comparisons on non-Latin-script targets therefore confound script with translation quality unless the script component is measured. We advocate publishing SBI, IPI, and Computational Tax alongside every such COMET score, turning an inherited assumption into a verifiable claim. The deeper fix is encoder-level: tokeniser and pre-training choices that distribute capacity more evenly across scripts. We plan to unify the three diagnostics into a single script-aware companion score in future work.

Limitations

Our analysis covers one encoder family (XLM-RoBERTa) and one metric (COMET); generalisation to other encoder–metric combinations requires further data. Our 1,400 sentences per language are smaller than the reference corpora in Brahma et al. (2025) and Kumar et al. (2017); we claim only that our dependent-vowel rank ordering is consistent with theirs and provides corroborating rather than independent evidence. A larger MQM-annotated dataset spanning many languages and scripts would enable a more robust analysis; to the best of our knowledge, no such dataset currently exists, and the WMT24 ENG-DEU and ENG-SPA release used as Latin-script controls covers only two language pairs. MAR is an instructive counter-case: mean COMET rises under romanisation ($+1.02$ points) while every quality indicator worsens, indicating that mean-level SBI normalisation alone is insufficient and severity-stratified correction is also required.

Artefact licences. All artefacts used—IndicMT Eval (Sai et al., 2023), WMT24 evaluation packages (Kocmi et al., 2024), XLM-RoBERTa, BLOOM-560m, COMET, BLEURT, BERTScore, and IndicXlit (Madhani et al., 2023)—are released for research use under permissive or open-research licences. Our analysis use is consistent with all terms; we do not redistribute the underlying data. Released code is MIT-licensed.

Ethics Statement

We foresee no ethical issues with this work. The study conducts systematic experiments to expose and quantify evaluation distortions caused by tokenisation artefacts in Indic scripts, offering interpretable statistical evidence and actionable recommendations. All data and benchmarks used are publicly available, and all released artefacts are analytical in nature with no foreseeable risk of harm. The analysis directly benefits under-served Indic-language NLP communities by surfacing previously hidden evaluation distortion and guiding more reliable metric usage.

References

Catherine Arnett and Benjamin K. Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19–24, 2025*, pages 6607–6623. Association for Computational Linguistics.

Maharaj Brahma, N J Karthika, Atul Singh, Devaraj Adiga, Smruti Bhate, Ganesh Ramakrishnan, Rohit Saluja, and Maunendra Sankar Desarkar. 2025. MorphTok: Morphologically grounded tokenization for Indic languages. In *Proceedings of the ICML 2025 Tokenization Workshop (TokShop)*, volume 267 of *Proceedings of Machine Learning Research*, Vancouver, Canada. PMLR.

Tanmay Chavan, Omkar Gokhale, Aditya Kane, Shantanu Patankar, and Raviraj Joshi. 2023. My boli: Code-mixed Marathi-English corpora, pretrained language models and evaluation benchmarks. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 242–249, Nusa Dua, Bali. Association for Computational Linguistics.

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, ACL 2020, Online, July 5–10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024, Torino, Italia, May 20–25, 2024*, pages 3553–3565. ELRA and ICCL.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Vitalii Hirak, Jaap Jumelet, and Arianna Bisazza. 2026. Assessing the impact of typological features on multilingual machine translation in the age of large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), EACL 2026, Rabat, Morocco*, pages 2416–2434. Association for Computational Linguistics.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and representation biases in multilingual models on dialectal NLP tasks. In *Proceedings of the 2025 Conference on*

Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4–9, 2025, pages 23992–24010. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7–11, 2022*, pages 9540–9561. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, Florida, USA, November 15–16, 2024*, pages 1–46. Association for Computational Linguistics.

Arun Kumar, Ryan Cotterell, Lluís Padró, and Antoni Oliver. 2017. Morphological analysis of the Dravidian language family. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017 (Volume 2: Short Papers)*, pages 217–222. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, (12):455–463.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh 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, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Aleksandar Petrov, Emilio La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, volume 36.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023*, pages 14210–14228. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Alexander Tsvetkov and Alon Kipnis. 2024. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12–16, 2024*, pages 7971–7989. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, and 1 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced african languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024, Mexico City, Mexico, June 16–21, 2024*, pages 5997–6023. Association for Computational Linguistics.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A Statistical Methods and Sample Sizes

The headline ANOVA on COMET groups target languages by script family (Devanagari: HIN, MAR; non-Devanagari: GUJ, MAL, TAM). Sample sizes and MQM severity bucket counts per language are in Table 5. ANOVA results, SBI threshold calibration, and the HIN–GUJ natural experiment are in Table 6.

Lang.	N	Def	VL	L	M	H	VH
GUJ	1,400	248	217	326	390	184	35
HIN	1,400	399	1	7	134	324	535
MAL	1,400	372	25	162	443	334	64
MAR	1,400	142	69	86	184	258	661
TAM	1,400	60	124	151	323	313	429

Table 5: Per-language sample sizes and MQM severity bucket counts. **Def** = Default (error-free); VL/L/M/H/VH = Very Low, Low, Medium, High, Very High.

Quantity	Native	Romanised
F	2,080.8	108.3
p	$< 10^{-99}$	3.5×10^{-25}
η^2	22.9%	1.5%
Sentence-level SBI ≥ 3.0 *proportions*		
GUJ	62.1%	
TAM	34.9%	
MAL	30.9%	
MAR	28.2%	
HIN	6.6%	
HIN–GUJ gap (Welch's t)		
Native	13.95 pts; $t = 41.7^{+}$, $p < 10^{-99}$	
Romanised	2.29 pts; $t = 6.84^{+}$, $p = 8.7 \times 10^{-12}$	

Table 6: ANOVA on COMET (script-family grouping, $N = 6{,}995$), sentence-level SBI thresholds, and the HIN–GUJ natural experiment. $\eta^2 = \mathrm{SS}_{\text{between}}/\mathrm{SS}_{\text{total}}$.

B Master Language Profile

Table 7 reports IP, TP, SBI, IPI, MATTR, and dependent-vowel rates for all seven language–script settings.

Lang.	Script	Family	$\overline{\mathrm{IP}}_{\mathrm{nat}}$	$\overline{\mathrm{TP}}_{\mathrm{nat}}$	$\mathrm{SBI}_{\mathrm{nat}}$	$\mathrm{IPI}_{\mathrm{nat}}$	$\mathrm{IPI}_{\mathrm{rom}}$	MATTR	DV/1k	IP–COMET r
SPA	Latin	Romance	1.009	1.090	1.12	0.009	—	0.581	—	-0.021^{-}
DEU	Latin	Germanic	0.535	1.133	2.24	0.465	—	0.634	—	-0.262^{+}
HIN	Devanagari	IA	0.636	1.210	2.03	0.364	0.735	0.633	18.3	$+0.203^{+}$
MAR	Devanagari	IA	0.490	1.165	2.61	0.510	0.720	0.827	24.5	-0.216^{+}
MAL	Malayalam	Dravidian	0.549	1.248	2.62	0.451	0.803	0.897	42.5	-0.266^{+}
TAM	Tamil	Dravidian	0.545	1.319	2.73	0.455	0.826	0.862	13.6	-0.263^{+}
GUJ	Gujarati	IA	0.438	1.388	3.45	0.562	0.711	0.752	36.7	-0.064^{-}

Table 7: Master language profile. All Indic values: Translation column, IndicMT Eval (Sai et al., 2023) ($N = 1{,}400$/language). ENG-DEU / ENG-SPA: WMT24 (Kocmi et al., 2024) (339 source sentences, pooled across MT systems: 6,006 rows for ENG-DEU and 4,651 for ENG-SPA). Ordered by $\mathrm{SBI}_{\mathrm{nat}}$ ascending. DV/1k is isolated dependent-vowel tokens per 1,000 XLM-R tokens, counted from tokeniser output. IA = Indo-Aryan. $^{+}p < 0.05$; $^{-}$not significant.

C Computational Tax Decomposition

Table 8 decomposes the Computational Tax into its Length Penalty (LP) and Entropy Penalty (EP) components for each language.

D Dependent-Vowel Isolation and Word Length

In Indian abugida scripts, vowels are written as diacritics (*dependent vowels*, DV) attached to the preceding consonant; a well-designed tokeniser preserves each consonant–vowel unit as a single token. XLM-RoBERTa's BPE vocabulary, trained on Latin-dominant text, frequently isolates these diacritics as standalone tokens carrying no semantic content. MAL produces 42.5 isolated DV tokens per 1,000 XLM-R tokens, GUJ produces 36.7, affecting 75.8% and 74.6% of sentences respectively (Table 7). TAM's DV rate is the lowest (13.6/1,000) yet TAM has the highest TP inflation under romanisation (+77.9%), a dissociation that points to a second, independent mechanism: TAM's agglutinative suffix chains produce long word forms BPE must split heavily, confirmed by TAM's word-length-to-TP Spearman $\rho = 0.921^{+}$ under romanisation (Table 9). MATTR, a vocabulary-free measure decoupled from XLM-R, reproduces the TP/IP hierarchy in the same order (MAL > TAM > MAR > GUJ > HIN), confirming morphological diversity (Hirak et al., 2026), not training-corpus artefacts.

Lang.	LP	EP	Tax	EP%
TAM	1.78×	3.13×	5.57×	66.4%
MAL	1.48×	2.78×	4.12×	72.2%
HIN	1.41×	2.40×	3.37×	71.9%
MAR	1.41×	1.75×	2.47×	62.1%
GUJ	1.15×	1.52×	1.75×	74.6%

Table 8: Computational Tax decomposition. $\mathrm{EP\%} = \ln(\mathrm{EP})/\ln(\mathrm{Tax})$. The Entropy Penalty dominates in every language; information loss per token, not token-count inflation, is the principal cost of romanisation.

Lang.	Native ρ	Romanised ρ
HIN	0.897^{+}	0.956^{+}
GUJ	0.883^{+}	0.952^{+}
MAR	0.891^{+}	0.942^{+}
TAM	0.850^{+}	0.905^{+}
MAL	0.875^{+}	0.900^{+}

Table 9: Sentence-level Spearman ρ between word count and XLM-R token count (all $p < 10^{-99}$, $N = 1{,}400$). Romanisation pushes correlations toward 1.0, the signature of single-character BPE fallback.

E Latin-Script Controls (DEU, SPA)

We use WMT24 ENG-DEU and ENG-SPA (Kocmi et al., 2024) as Latin-script controls. SPA anchors the Parity zone (IPI = 0.009); DEU provides the critical refutation of the script-surface explanation. DEU uses the same Latin alphabet as SPA yet exhibits IP suppression of the same magnitude as the Indic native-script range (0.535 vs. Indic 0.438 to 0.636). The IP–COMET Pearson correlations diverge by a factor of twelve ($r = -0.262^{+}$ for DEU vs. $r = -0.021^{-}$ for SPA; $N \approx 28{,}000$). Byte premiums are near-identical (1.25× vs. 1.09×), directly ruling out character-encoding overhead.

DEU's productive compound-noun morphology drives the gap: e.g., *Erlassvereinfachungsverträge* (28 chars, "decree simplification contracts") $\rightarrow$ 9 XLM-R tokens, versus *simultáneamente* (15 chars, "simultaneously") $\rightarrow$ 4 tokens (Ex-D). Only 14.6% of DEU word types receive single-token XLM-R encoding versus 22.1% for SPA; long DEU words require 32% more tokens than long SPA words. DEU MATTR (0.634) exceeds SPA (0.581), confirming greater morphological surface diversity consistent with the compound-noun mechanism.

The SBI Reversal. In Indic languages, script-induced SBI elevation is quality-independent: fragmentation operates as a fixed structural offset applied equally across translations. In DEU, erroneous MT outputs tend to be paraphrastic or syntactically simplified (Zouhar et al., 2024), producing shorter, more common word sequences and therefore *lower* fragmentation. Error-free DEU sentences yield mean $\mathrm{SBI} = 2.349$ whereas major-error sentences yield $\mathrm{SBI} = 2.129$ (Kruskal–Wallis $H = 135.0^{+}$, $p = 5.0{\times}10^{-30}$), the opposite of the romanised-Indic direction. SBI remains a valid zone diagnostic in both regimes: it correctly places DEU in Burden ($\mathrm{SBI} = 2.24$) and SPA in Parity ($\mathrm{SBI} = 1.12$).

Setting	$\overline{\mathrm{IP}}$	SBI	MATTR	Byte pr.	1-tok%	IP–COMET r
ENG-SPA (SPA)	1.009	1.12	0.581	1.09×	22.1%	-0.021^{-}
ENG-DEU (DEU)	0.535	2.24	0.634	1.25×	14.6%	-0.262^{+}
Indic nat.	0.438–0.636	2.03–3.45	0.633–0.897	2.45–5.33×	8.7–30.3%	varies

Table 10: Latin-script controls (WMT24) alongside the Indic native range. **1-tok%** is the fraction of unique word types receiving single-token XLM-R encoding. IPI zones: ENG-SPA = Parity; ENG-DEU = Burden; Indic native = Burden; Indic romanised = Paradox.

F Corpus-Level Corroboration

Table 11 reports MATTR and Byte Premium across all seven settings as vocabulary-free and encoding-independent sanity checks on the TP/IP findings.

Lang.	Script	$\mathrm{MATTR}_{\mathrm{nat}}$	$\mathrm{MATTR}_{\mathrm{rom}}$	Byte Prem.
HIN	Devanagari	0.633	0.601↓	2.45
MAR	Devanagari	0.827	0.721↓	3.64
GUJ	Gujarati	0.752	0.665↓	3.03
TAM	Tamil	0.862	0.733↓	5.30
MAL	Malayalam	0.897	0.768↓	5.33
SPA	Latin (Romance)	0.581	—	1.09
DEU	Latin (Germanic)	0.634	—	1.25

Table 11: Corpus-level corroboration across all seven settings, ordered by SBI. MATTR: $w = 500$, whitespace tokenisation (Covington and McFall, 2010). Byte Premium: UTF-8 bytes-per-word ratio, target to English (Arnett and Bergen, 2025). ↓ = falls under romanisation.

G Axiomatic Justification of the Diagnostics

We give a brief axiomatic motivation for the three novel diagnostics.

Why SBI is a ratio. A diagnostic of representational burden should: (a) equal 1.0 when $\mathrm{TP} = \mathrm{IP} = 1.0$; (b) be insensitive to constant rescaling of either TP or IP; (c) diverge as either fragmentation grows or information density collapses; (d) recognise the Gujarati anomaly, where small TP inflation combines with very low IP to yield high overall burden. Additive $\mathrm{TP} - \mathrm{IP}$ fails (a) and (d). Ratio TP/IP is the unique simplest function satisfying all four.

Why Tax is multiplicative. The romanisation overhead has two components paid simultaneously: a sentence takes more tokens to encode (LP) and conveys less per token (EP). A practitioner cannot trade one penalty for the other; their joint effect on inference cost is multiplicative, not additive. TAM $\mathrm{LP} = 1.78, \mathrm{EP} = 3.13$ is categorically different from $\mathrm{LP} = 3.13, \mathrm{EP} = 1.78$ despite identical additive sums.

IPI threshold calibration. The Paradox boundary at $\mathrm{IPI} = 0.70$ is empirically calibrated. Romanised $\rho \in \{0.230, 0.240, 0.278, 0.341, 0.460\}$ for the five Indic languages; four of five fall below $\rho = 0.40$, with MAL ($\rho = 0.460$) the sole exception (still a -29.3% drop from native $\rho = 0.651$). The Burden upper bound at $\mathrm{IPI} = 0.70$ is the maximal IPI consistent with $\rho \geq 0.40$ across all native-script conditions.