Which Metric for Which Script? A Quantified Meta-Evaluation of Automatic Metrics and LLM Judges for Low-Resource Indic Machine Translation

Rudali Huidrom[1]†, Vikas Kumar[2]†, Hoomexsun Pangsatabam[3], Pinaki Das[4,5], Kshetrimayum Boynao Singh[6,7], Anand Konjengbam[2]
[1]IIIT Delhi [2]IIT Dharwad [3]NIT Manipur [4]IIT Bombay [5]Monash University
[6] NIT Silchar [7] IIT Patna
Correspondence: huidromrudali@gmail.com

Abstract

Machine translation shared tasks publish several automatic metrics and rank systems on one of them, assuming the choice is inconsequential. We test this assumption on the WMT 2025 and WMT 2026 Low-Resource Indic Language Translation tasks, where pre-trained multilingual encoders are unavailable for several target scripts and languages. As no human gold standard evaluation exists at scale, we adapt the Quantified Reproducibility Assessment (QRA) framework across 20 WMT26 and 14 WMT25 directions, to measure agreement between six automatic metrics and two LLM judges as independent measuring instruments. We show that QRA's Type I measure fails under this repurposing when instruments sit on different scales, and suggest a rank-based repair. Using a script-controlled natural experiment (Manipuri evaluated in both Bengali and Meitei Mayek) and a null-output validity test, we observe that neural metrics inflate scores for text in scripts outside the encoder's coverage and compress it into a narrow discriminative band. Switching to Meitei Mayek raises BERTScore and COMET by 21.6 and 23.3 points, respectively, though language and content are unchanged, and no metric preserves system ranking; on degenerate output the two retain 0.81 and 0.61 points on English-to-Indic directions. The effect reverses in the opposite direction. Metric choice should follow script coverage, not language identity. Lacking human judgments, our claims concern instrument behaviour, not human preference. The code and resources are available on our Github.[2]

† The first two authors made equal contributions to all aspects of the main paper; both are designated as first authors.

[2]https://github.com/RHuidrom96/which-metric-which-script.git

1 Introduction

Machine translation shared tasks have converged on a reporting convention: publish a table of automatic metrics, and rank participating systems on one of them. The Workshop on Machine Translation (WMT) 2025 Low-Resource Indic Language Translation task (Pakray et al., 2025, 2026) follows this convention. It reports BLEU, METEOR, TER, chrF++, BERTScore, ROUGE-L, chrF, Cosine-Similarity and COMET. WMT 2025 covered 7 language pairs, spanning 14 translation directions: Assamese, Manipuri, Khasi, Mizo, Nyishi, Bodo, and Kokborok, each paired with English. WMT 2026 covered 10 language pairs, spanning 20 translation directions: Assamese, Mizo, Khasi, Manipuri (in Bengali and Meitei Mayek scripts), Bodo, Kokborok, Karbi, Nagamese, and Tagin. Nyishi was withdrawn from the 2026 edition.

The convention carries an implicit assumption: that the metrics broadly agree, so that the choice of which to rank on is a matter of taste rather than of consequence. For high-resource language pairs, this assumption is defensible and has been extensively validated against human judgments (Freitag et al., 2021, 2022, 2023). For the languages in this task, it has never been tested. Several of them are written in scripts that appear in no major multilingual encoder's pretraining data; several have no meaningful presence in the corpora on which COMET and BERTScore were built. A metric whose backbone cannot represent the target language at all will still return a number, which will still be used to rank systems.

This paper asks what those numbers are worth. We take the WMT25 & WMT26 results table as our object of study and treat metric disagreement as the phenomenon to

Proceedings of the Eleventh Conference on Machine Translation, pages 281–303
October 28-29, 2026 ©2026 Association for Computational Linguistics

be measured rather than as noise to be averaged away. Meta-evaluation of MT metrics is conventionally framed in terms of correlation with human judgment. That framing requires a gold-standard, but for the languages under consideration, there is no gold-standard judgment at scale. We therefore adopt a different apparatus: the Quantified Reproducibility Assessment framework of Belz et al. (2026), which supplies a taxonomy of four result types and a matching set of continuous-valued measures reported at system, quality-criterion and study level. QRA is designed to quantify agreement between repeat measurements. We repurpose it to quantify agreement between instruments, which the International Vocabulary of Metrology explicitly licenses[3]: §3.2 lists different measuring systems we used: six metrics and two LLM judges become eight measuring systems applied to a common measurand. QRA's Type I measure, the bias-corrected coefficient of variation (CV*), presupposes that the values being compared share a scale. Metrics do not: on WMT26, BERTScore occupies an interquartile range of [81.9, 94.5] while BLEU occupies [7.4, 22.4]; on WMT25, the offset is starker still, with METEOR reported on a 0–1 scale (IQR [0.08, 0.45]) alongside BLEU on 0–100 (IQR [2.1, 13.7]). Applied naively, CV* reports catastrophic disagreement that is entirely an artefact of scale offset. We report this as a negative result, and propose a rank-based system-level measure that answers a related but distinct question.

The task design contains an unusually clean identification strategy. Manipuri is evaluated twice in WMT26, once in Bengali script and once in Meitei Mayek, with nine teams submitting to both English→Manipuri directions and eight to both Manipuri→English directions. Language, content, reference construction and system are held constant; only the orthography of the target varies. A metric that measures translation quality should be approximately invariant to this change. A metric that is in fact measuring the tractability of the target string to its own tokeniser should not be. The remaining languages supply the converse contrast: six WMT26 targets are written in Latin script and two in Bengali, so script is held constant while pretraining coverage varies. Manipuri isolates orthography at fixed language; these sets isolate language at fixed orthography. We formalise this as a coverage-tier analysis in §4.3.

A natural experiment can show that metrics disagree, and can show which of them behaves implausibly, but it cannot by itself say which is right when two metrics diverge on a specific segment. For that we turn to LLM-as-judge, applied not as a substitute gold standard but as an adjudication instrument on a targeted sample: the segments where COMET is high and BLEU is low. Such segments have two possible explanations. Either the translation is adequate but lexically, morphologically or syntactically distant from the reference, in which case BLEU is penalising legitimate variation and COMET is correct; or the translation is fluent but unfaithful, in which case COMET is rewarding surface plausibility and BLEU is correct. Distinguishing these is a well-posed question that a judge model can answer at the segment level, and it is the question our error taxonomy is built to separate.[4]

Our contributions are as follows:

- A script-controlled natural experiment showing that BERTScore and COMET inflate by over 20 points when the same language is written in a script outside the encoder's coverage, and that no metric preserves system ranking across that change.
- The first application of the QRA Type I, II and IV apparatus to an MT shared task, reported at all three of its levels.
- A documented failure of QRA Type I under cross-instrument comparison, together with a rank-based repair and a size-controlled variant.
- Evidence on discriminative power and floor behaviour: BERTScore compresses into a 12.6-point band in WMT26 while WMT25's embedding-based Cosine-Similarity is the most dispersed instrument of that year at 62.4% of its scale,

[3] https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf

[4] We selected our LLM-as-Judge models based on the translation accuracy of the low-resource language pair (Mni_Beng→Eng) (BLEU scores; *Gemini 3.1 Flash Lite*: 28.18, and *Claude Opus 4.8*: 27.17). Claude Opus 4.8 was the version available at selection time; Opus 5 was used for the judge runs reported below.

and the scores that all six metrics assign to output known a priori to be worthless.

- An LLM-judge adjudication of BLEU–COMET disagreement cases on English-Manipuri, distinguishing reference mismatch from genuine translation error.
- Tier-conditional recommendations for metric selection and shared-task reporting.

2 Related Work

2.1 Automatic metrics

MT evaluation splits into two lines. Surface metrics score n-gram overlap with a reference: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF/chrF++ (Popović, 2015, 2017). They need no target-language model, but correlate weakly with quality once systems are close (Mathur et al., 2020). Neural metrics instead score similarity in a pretrained encoder's embedding space: BERTScore (Zhang* et al., 2020) and COMET (Rei et al., 2022). WMT22 found these more reliable than BLEU and recommended dropping it (Freitag et al., 2022). Both rely on XLM-R-style encoders (Conneau et al., 2020), whose pretraining coverage is uneven across languages and scripts.

2.2 Meta-evaluation

WMT's Metrics Shared Task ranks metrics by correlation with expert human ratings (Freitag et al., 2021, 2022, 2023). This needs a large annotated pool, which does not exist for most languages studied here. Mathur et al. (2020) showed that correlations with human judgments depend strongly on the set of evaluated systems, limiting the usefulness of raw correlation scores alone. Quantified Reproducibility Assessment (QRA) offers an alternative by evaluating automatic metrics as measurement instruments rather than against a human gold standard. It quantifies reproducibility using metrology-inspired measures of agreement and variability between evaluation instruments (Belz et al., 2022; Belz, 2022). QRA++ extends this to a four-type taxonomy of NLP results (Belz, 2025). Both were built for repeated measurement by a single method, not for comparing structurally different instruments applied once each. No prior study has applied the QRA++ apparatus to the full metric and judgment suite of a shared task.

2.3 Coverage and script bias

Pretraining coverage is uneven: XLM-R performance generally correlates with the amount of pretraining data available for each language (Conneau et al., 2020). Tokeniser studies report a similar disparity at the subword level: under-represented languages and scripts often require substantially more tokens to encode the same content (Petrov et al., 2023), increasing inference costs for multilingual language models. This work explains why coverage should distort scores, but it does not measure that distortion for a metric under controlled conditions.

2.4 Low-resource Indic MT

NLLB (Team et al., 2022), IndicTrans2 (Gala et al., 2023) and Sarvam-translate (Sarvam AI, 2025) cover most Indic languages studied here, and shared-task systems fine-tune similar backbones. These papers report BLEU, chrF, and COMET-style scores as evidence of quality, without testing whether the metric itself is valid for the language-script pair being scored.

2.5 LLM-as-judge

A newer line of work scores translations by prompting a general-purpose language model rather than training a dedicated metric. GEMBA shows that zero-shot GPT prompting can achieve competitive, state-of-the-art system-level accuracy against human MQM judgments (Kocmi and Federmann, 2023). Zheng et al. (2023) study LLM judges more broadly and reports position bias, verbosity bias, and self-preference bias. These biases carry over to translation judging unless checked directly. The models used as judges, including GPT-4 (OpenAI, 2023), Gemini (Gemini Team, 2023), and Claude (Anthropic, 2024), are trained on the same skewed script distribution as the embedding encoders behind neural metrics. Whether this skew affects judge scores on low-coverage scripts remains untested.

3 Dataset and Instruments

3.1 Data Used

We use the official results of the WMT Low-Resource Indic Language Translation shared

task, WMT25 (Pakray et al., 2025) and WMT26 (Pakray et al., 2026). WMT25 covers 14 directions across 7 languages (Assamese, Manipuri, Khasi, Mizo, Nyishi, Bodo, Kokborok). It has 18 participating teams and up to 8 primary submissions per direction. Scores are reported on BLEU, METEOR, ROUGE-L, chrF, TER, and Cosine-Similarity, the last restricted to Indic→English directions. WMT26 extends the task to 20 directions across 10 languages. Nyishi is withdrawn; Karbi, Nagamese, and Tagin are added, and Manipuri is split into Bengali and Meitei-Mayek script variants (Huidrom et al., 2026). It has 24 participating teams and up to 15 primary submissions per direction. Its metric suite is BLEU, METEOR, TER, chrF++, BERTScore, and COMET, replacing ROUGE-L/Cosine-Similarity with two embedding-based metrics. Both editions report only system-level scores, split into primary and contrastive/constrained submissions per team. We experiment and report our results and analyses on the languages mentioned above.

As a special case, we choose a specific subset from the WMT 2025 test set (Pakray et al., 2025), specifically the directions of translation from and into Manipuri to English (mni→en and en→mni), in order to diagnose errors at the segment level as well as LLM-as-a-judge experiments. The only occurrence of the Manipuri language in the WMT 2025 benchmark is in the Bengali writing system. To study the effects of metric divergence along with subword segmentation and model architectures under this low-resource condition, we use the translation outputs of our competing submission pipeline, which follows the iterative back-translation and fine-tuning methodology (bt+ft+iter)[5] using subword regularization atop the IndicTrans2 multilingual backbone. Using this controlled dataset allows us to directly test LLM evaluators against surface and neural metric divergence on a real-world, highly agglutinative target language.

3.2 Instruments Used

The instruments used in our study can be classified into three categories:

- **Surface N-gram Metrics**: In WMT25, n-gram metrics include BLEU, METEOR, ROUGE-L, chrF, and TER, while WMT26 includes BLEU, METEOR, TER, and chrF++. These metrics measure lexical and character-level overlap with a reference string.
- **Neural Embedding Metrics**: These metrics calculate similarity in the embedding space of a pre-trained multilingual encoder. It includes metrics such as Cosine-Similarity (WMT25), BERTScore and COMET (WMT26).
- **LLM Judges**: For our segment-level diagnostics and error adjudication, we employ two cutting-edge large language models as evaluators: *claude-opus-5* and *gemini-3.1-flash-lite*.

4 Experimental Design

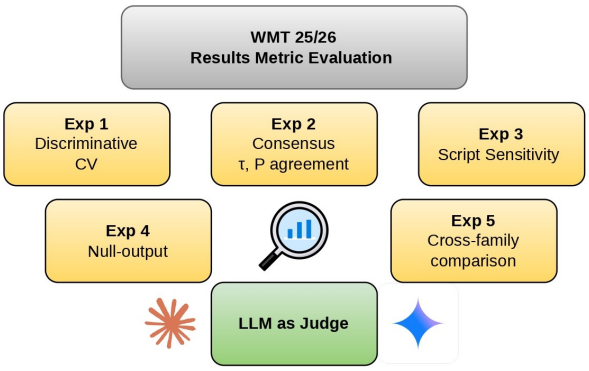


Figure 1: Overview of WMT 2025/2026 results metric evaluation experiments.

We report five experiments, ordered by the strength of the claim each can support as presented in Figure 1. Experiments 1 (§4.1) and 2 (§4.2) establish what the instruments do relative to one another: whether each can separate systems at all (Type I), and how far each agrees with the consensus of the rest (Types II and IV). Neither licenses a validity claim, since instruments may agree and be jointly wrong. Experiments 3 (§4.3) and 4 (§4.4) supply that validity anchor without human annotation, by constructing conditions in which the correct answer is known in advance, a script change that should leave scores invariant, and degenerate output that should score near zero. Experiment 5 (§4.5) then adjudicates the residual: for segments where two metrics diverge

[5]bt: back translation, ft: forward translation, iter: iterative pseudo labelling

and neither prior construction applies, we ask a second instrument family which of them is right.

4.1 Experiment 1: Discriminative power (Type I over systems)

An instrument with a negligible spread across systems cannot support a ranking, however well it correlates with anything else. Experiment 1 therefore precedes all correlation analysis: it asks whether each instrument separates the objects of measurement at all.

For each language pair we compute CV* of each instrument across systems, together with the observed range and the effective number of distinct values, aggregated at the quality-criterion and study levels. Alongside this we report the rank-based Type I of §A.2 in the QRA3 template at the system, quality-criterion and study level, together with its size-controlled variant, in which each pair is subsampled to a fixed K systems over B draws (§A.2, Size Control; §A.3) so that dispersion is comparable across pairs of unequal size. For the system-type analysis supporting Narrative (b), CV* over instruments is additionally grouped by the taxonomy, quantifying how far a system type's standing depends on which instrument is chosen.

TER is unbounded above, and its cross-system CV* is inflated accordingly; we report it with that caveat rather than reading it as discriminative strength.

4.2 Experiment 2: Consensus centrality (Types II and IV)

Experiment 2 asks which instruments agree with the consensus of the others, and where that agreement breaks down.

For each language pair we compute all pairwise r, ρ and τ_b between instruments, Kendall's W over the full instrument set, and P over all instrument pairs (§A.1 Measures). A leave-one-out summary reports, for each instrument, its mean τ_b and mean P against the remaining instruments. Results are given at the quality-criterion and study level (Table 4), and computed separately for metrics alone and for metrics together with the judge scores. We additionally test whether concordance varies with task difficulty, by regressing quality-criterion-level W on mean BLEU, and with the number and diversity of entrants.

We report r, ρ and τ_b jointly rather than selecting one, following the rationale of Belz and Thomson (2026) that the three afford different perspectives on the same set of scores. This is not a formality on these data: we observe language pairs in which a single outlier system sustains a near-perfect Pearson correlation between two instruments whose rank agreement is approximately zero from this experiment.

4.3 Experiment 3: Script sensitivity and coverage tiers

Experiment 3 is the paper's identification strategy. It asks whether instrument scores are driven by translation quality or by whether the target script falls within the instrument's pretraining coverage, and it does so by exploiting a natural experiment embedded in the task design.

Manipuri is evaluated in two scripts, Bengali and Meitei Mayek, with nine (En→Mni) and eight (Mni→En) systems submitted per script. Since only orthography varies across these paired subsets, a valid metric should score them equally; deviation signals an instrument artefact, not a quality difference, giving a validity test that requires no human annotation.

For each paired system we compute the per-instrument score shift, Meitei Mayek minus Bengali, with bootstrap intervals on the mean shift (Table 5), and for each instrument the Kendall τ_b between its rankings of the paired systems under the two scripts. The Manipuri-to-English direction serves as a control: the output language is English in both conditions, so shifts should be markedly smaller there than in the English-to-Manipuri direction. An instrument that fails this control is unstable for reasons beyond target-script coverage, and we report such failures separately. §4.2 and §6.1 provide insights into the data used and the impact of script coverage bias, particularly for the Manipuri language, including a discussion of any expected or significant discrepancies related to script coverage.

4.4 Experiment 4: Null-output validity testing

Experiment 4 supplies a second validity anchor from a second class of cases in which the cor-

rect answer is known *a priori*. Where Experiment 3 fixes quality and varies script, Experiment 4 fixes the verdict, the output is worthless, and asks what each instrument returns. More details in Appendices D and E.

We identify degenerate primary submissions, defined as those scoring BLEU ≤ 1.0, a set that includes empty outputs, wrong-script outputs and outputs in the wrong language. For each instrument we report its floor over this set and the *floor-inflation ratio*: the instrument's score on the degenerate system divided by its score on the top-ranked system in the same language pair, with bootstrap intervals (Table 6). The analysis is framed as a one-sided test against the null that the instrument score is independent of the degenerate-competent distinction.

An instrument whose floor sits close to its ceiling cannot support absolute quality claims even where its rankings remain serviceable. We report this distinction of rank-preserving but not scale-meaningful explicitly in this paper.

4.5 Experiment 5: Cross-family comparison and adjudication

Experiments 3 and 4 establish that certain instruments behave implausibly under constructed conditions. They do not determine which instrument is correct on an ordinary segment where two diverge. Experiment 5 addresses that question on a targeted sample, using the judge protocol of §4.6.

We compute the Type II and Type IV apparatus across instrument families, as set out in Table 7. Between the n judges we report Kendall's τ_b and P on each scalar criterion at segment level; with $k = 2$, Kendall's W reduces to a function of Spearman's ρ, which we report in its place. Cohen's κ between judges on the error-type label constitutes the only Type III (§A.1, Measures) result in the paper. Type I is computed as CV* between judges on each scalar criterion, per segment and aggregated to criterion and study level, reported with bootstrap intervals given the noise inherent at $n = 2$.

4.6 LLM-as-judge protocol

To better understand the divergence between COMET and n-gram overlap-based metrics such as BLEU, we conduct a qualitative analysis using an *LLM-as-a-Judge* framework. Specifically, we evaluate individual test examples to characterize the nature of the translations produced by our systems. The LLM is prompted to assess each translation along three dimensions: *adequacy*, *fluency*, and *semantic equivalence* with respect to the reference translation. Beyond these evaluation criteria, the judge is also instructed to identify and categorize *errors* in the system output into five predefined classes: *lexical variation*, *word-order variation*, *morphological variation*, *named-entity variation*, and *genuine meaning error*. For each identified error, the LLM additionally provides a *confidence* score and a *textual rationale* explaining the assigned error category. This analysis allows us to distinguish cases where metric discrepancies arise from linguistic variation from those that reflect genuine semantic errors.

We use *claude-opus-5* and *gemini-3.1-flash-lite* for our experiments. The prompts used for this experiments can be seen in appendix 21.

5 Results & Analysis

5.1 Experiment 1: Discriminative Power (CV*)

Table 3 shows a consistent pattern across both years: surface n-gram metrics (BLEU, ROUGE-L, METEOR) reach CV* (lower the better) of 44–74%, while embedding-based metrics compress this spread sharply. BERTScore and COMET fall to 6.5% and 17.4% in WMT2026, and Cosine-Similarity to 40.7% in WMT2025.

This gap widens further on the lowest-resource, script-uncovered pairs (e.g. English→Kokborok, BLEU CV* = 162.46%, n=2 systems), a pattern Experiment 3 examines directly by coverage tier. At the WMT2026 category level, Category 2 systems get a higher CV* than Category 1 for every instrument except TER, and BERTScore stays flat across categories, so its low discriminative power is not category-specific. Several WMT2025 pairs have only 2–3 systems, widening their Confidence Interval(CIs), so study-level numbers are more reliable than individual QC-level cells. Overall, BLEU-family metrics separate systems most in both years,

while BERTScore/COMET's low CV* raises the RQ1 concern that neural metrics may lack the spread needed for fine-grained ranking in this setting.

5.2 Experiment 2: Consensus Centrality

Table 4 shows TER as the least central instrument in both years, well behind the rest. In WMT2026, chrF++ is most central and every other instrument, including BLEU and the neural metrics, is significantly less central after Holm correction ($p_{\text{adj}} < 0.05$).

In WMT2025, Cosine-Similarity ranks highest, but with only 2–9 systems per pair (several pairs at just 2–3), none of the top four instruments differ significantly from one another.

Figure 2 confirms TER correlates weakest with every other instrument in both years (τ_b 0.59–0.65 in 2026, 0.63–0.76 in 2025), while WMT2026's chrF++/BERTScore/COMET and WMT2025's METEOR/chrF/Cosine-Similarity form tight clusters ($\tau_b \geq 0.85$). At the QC level, agreement drops on the lowest-resource pairs: WMT2026's English→Karbi (W=0.44) and English→Tagin (W=0.54) are weakest, and two WMT2025 pairs (Bodo→English, Kokborok→English) have too few systems for W/P to be computed. Instrument agreement is high overall (median $\tau_b > 0.8$), with TER as the only consistent outlier across both datasets.

5.3 Experiment 3: Script Sensitivity and Coverage Tiers

Table 5 shows that the script effect is significant only for some instruments and flips direction between the two translation directions. In en→indic, BERTScore and COMET inflate Meitei-Mayek scores over Bengali (both CIs exclude 0); the lexical metrics show no significant shift. In indic→en, this reverses: BLEU, METEOR, chrF++ and COMET score Meitei-Mayek significantly lower, and TER significantly worse; only BERTScore is not significant. No instrument is script-biased in the same direction both ways. Figure 3 shows top-median inflation rising with coverage tier[6,7] for TER, METEOR and chrF++ in WMT2026, while BERTScore stays flat and low (0.9–4.2) at every tier; COMET is flat through tier 3 (1.5–4.8) then jumps to 11.1 at tier 4. BLEU peaks at tier 3 (13.1) and falls at tier 4 (9.3). WMT2025 shows no clean monotonic trend: BLEU, TER and chrF peak at tier 2 and fall at tier 4, likely because those tier-4 pairs have only 2–3 primary systems, capping the top-median gap; WMT2025 also has no tier-3 pairs, since its Manipuri is undifferentiated and assigned to tier 2. METEOR, ROUGE-L and Cosine-Similarity stay near zero across all WMT2025 tiers, but this reflects their 0–1 scale, not genuine flatness. Overall, script and coverage effects are evident for several instruments but depend on direction and year, rather than following one consistent pattern.

5.4 Experiment 4: Null-Output Validity

We flag primary submissions with BLEU $\leq$ 1.0 as degenerate and compute each instrument's *floor-inflation ratio*: score on the degenerate system over score on the pair's top primary system (inverted for TER), so 0 is ideal and 1 is worst. WMT26 has 16 degenerate systems across 11/20 pairs; WMT25 has 5 across 4/14 pairs. BLEU and METEOR stay near the floor in both years (ratios $\leq$0.11, Table 6), confirming surface-overlap metrics are validity-safe on null output. The worst offender differs by year and is not a fixed metric-family property: WMT26's embedding-based metrics fail most, with BERTScore (0.81) and COMET (0.61) unable to separate degenerate from top-system output, concentrated on Tier 3–4 pairs such as Manipuri (Meitei-Mayek) and Tagin (Figure 4, also see §4.4 for more details), a coverage-driven failure; WMT25 has no embedding metric, and instead TER fails worst (0.42), driven almost entirely by a single short-reference case, English→Kokborok (0.99). Since the two years share no common failing instrument, null-output inflation is instrument- and dataset-specific, and Narrative (a)'s tier-conditional recommendation should weight per-pair ratios over the study-level mean.

[6] Tier defines how well a language/script pair is represented in the pretraining data of the encoders behind BERTScore and COMET.

[7] Tier 1 = well-covered by the multilingual encoder's pretraining; Tier 4 = essentially absent from it, with Tiers 2–3 as intermediate degrees of under-representation.

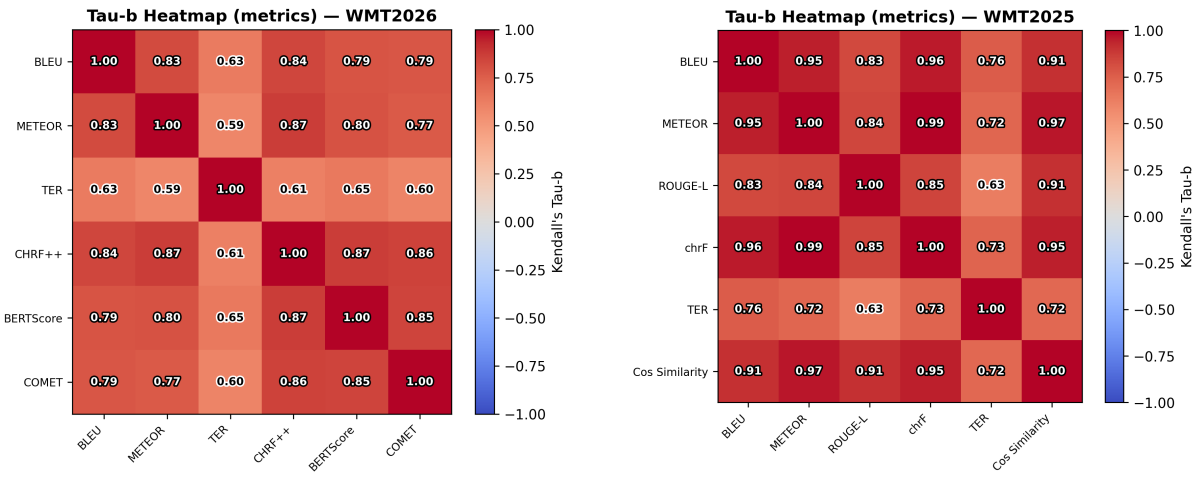


Figure 2: Instrument–instrument τ_b heatmap, metrics scope. Left: WMT2026. Right: WMT2025.

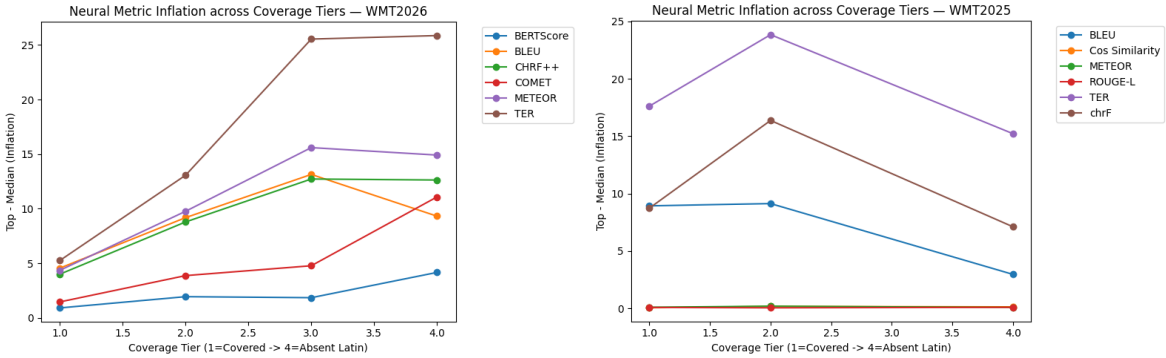


Figure 3: Top-minus-median inflation by coverage tier, all instruments overlaid. Left: WMT2026. Right: WMT2025.

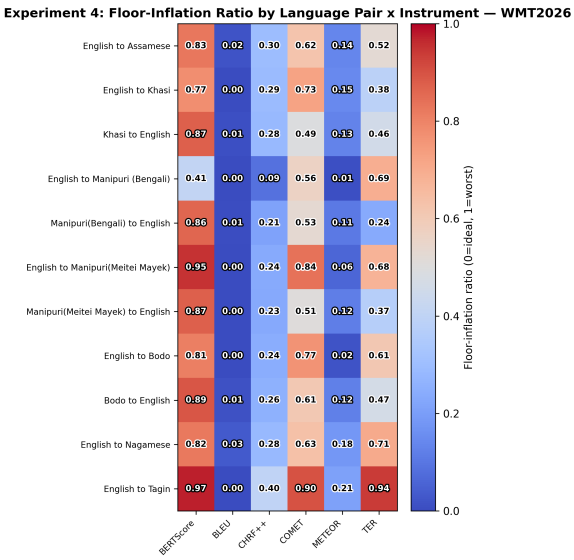
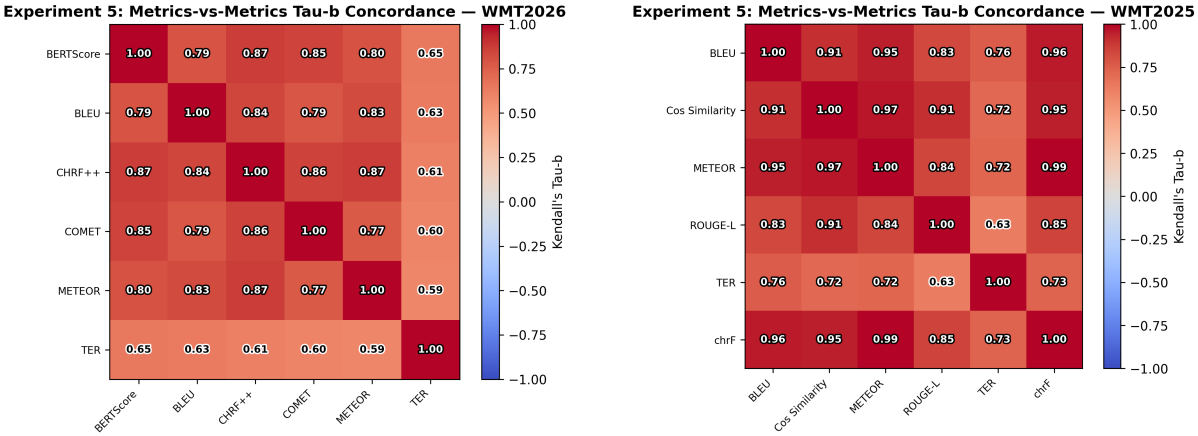


Figure 4: WMT26: floor-inflation ratio by language pair × instrument; blank cells indicate no degenerate submission for that pair.

5.5 Experiment 5: Metrics-vs-Metrics Concordance

We report internal concordance of the automatic-metric family only (Table 7: all pairwise instrument-instrument Kendall's τ-b, system level, aggregated across all language pairs; WMT26: 6 metrics, 15 instrument pairs, over 20 language pairs; WMT25: 6 metrics, 15 instrument pairs, over 14 language pairs). While figure 2 established which metrics form a general consensus, figure 5 analyses these pairwise correlations to demonstrate that a metric's underlying computational family does not dictate its alignment with others. In both years, metric concordance is generally high (τ-b mostly >0.75) but not uniform: TER is the least concordant instrument with every other metric in both releases (WMT26: τ-b 0.59–0.65, Figure 5a; WMT25: 0.63–0.76, Figure 5b), consistent with it being an edit-distance rather than n-gram-overlap or embedding-



(a) WMT26

(b) WMT25

Figure 5: instrument-instrument Kendall's τ-b concordance (metrics vs. metrics), system level.

based measure. The highest concordance in WMT26 reaches 0.87, tied between chrF++/BERTScore and chrF++/METEOR; in WMT25 it is METEOR/chrF (0.99), with Cosine-Similarity and the n-gram metrics (BLEU, METEOR, ROUGE-L, chrF) also mutually near 0.90–0.97. WMT26's embedding metrics (BERTScore, COMET) do not stand apart from the surface metrics here, both correlate with BLEU/METEOR/chrF++ at 0.77–0.87, comparable to the surface metrics' concordance with each other, so at the ranking-agreement level (unlike the floor-inflation results in §5.4) metric family does not predict internal disagreement; only TER consistently does.

5.6 LLM-as-a-Judge

We extract the 100 most divergent examples of BLEU v/s COMET and analyse them. We understand the following.

Root of Metric Divergence From Figure 6a, we observe that *Lexical Variations*, Word-Order Variations and *Morphological Variations* constitute the highest number of cases in our error categorizations. Naturally, these constitute the most to the divergence seen in the BLEU and COMET scores.

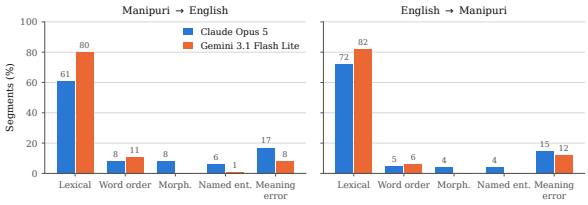
Judge Agreement and Metric Analysis In Figure 6b, we see *Claude* having more instances of categorizing into various error types, against *Gemini*, which often tries to club error types together. Meaning errors receive low BLEU scores. The judges agree on surface variation, and most such points lie above the diagonal in Figure 6c. Surface variation errors also hurt COMET and BLEU scores, and hence judges agree on them as well, with such points lying closer to the origin in general. Interestingly, we observe a larger spread of points, particularly in the upper half, for En→Mni in Figure 6c than for Mni→En. This suggests that for Manipuri translations, models struggle to match the references; however, they make up for it in the semantic space with high COMET scores.

6 Discussion

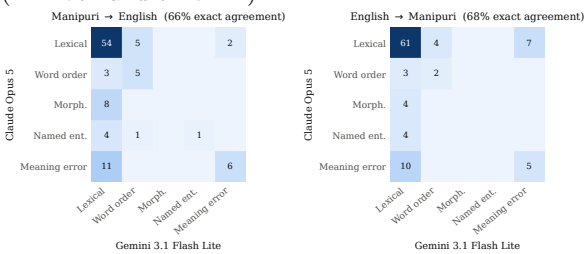
Our meta-evaluation shows that automatic metrics' assumptions diverge from their actual performance on low-resource Indic languages, making single-metric system ranking unreliable for non-pretrained scripts.

6.1 Impact of Script Coverage

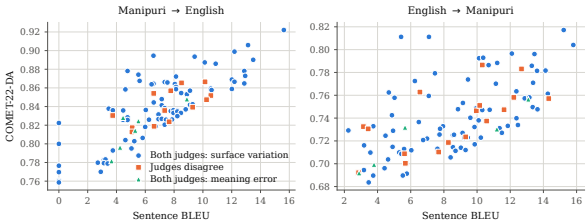
The findings of Experiment 1 and Experiment 3 clearly indicate that script coverage plays a critical role in determining the consistency of neural metrics. Metrics such as BERTScore and COMET, which are based on embeddings, suffer from score compression, making it difficult to distinguish between systems for low-resource language pairs. Neural metrics show inflated scores when evaluating texts written in uncovered scripts such as Meitei-Mayek (Manipuri) and Tier 4 Latin scripts. Additionally, it should be noted that the language script was observed by automatic measures that employ language models (such COMET,



(a) Distribution of error categories identified by Claude and Gemini judges across both translation directions (mni→en and en→mni).



(b) Confusion matrix showing agreement and disagreement between Claude and Gemini judges on error classification verdicts.



(c) COMET and BLEU scores stratified by joint judge verdict, illustrating the relationship between automatic metrics and judge-assigned error categories.

Figure 6: Error analysis across three dimensions: error category distribution, inter-judge agreement patterns, and metric-verdict alignment.

etc.). For instance, the Meitei Mayek script of Manipuri would not have been exposed because it is a standalone script, subjecting it to script coverage bias, whereas the Bengali script used in Manipuri would have been seen due to exposure to Bengali/Assamese language, even though Manipuri language may or may not have been seen by the language model itself. This is our theory, and we hope to investigate it more in the future.

6.2 Metric Validity on Null Output

Agreement of the instrument does not mean that the instrument is valid. Experiment 4 highlights an important weakness of embedding-based scores, which overrate the degenerate or nearly worthless systems. BERTScore and COMET were unable to distinguish top systems from degenerate systems in Tier 3-Tier 4 pairs. In contrast, overlap scores such as BLEU and METEOR rated the degenerate systems nearly at the floor level, showing their higher validity when assessing null systems.

6.3 Metric Divergence with LLM Judges

Despite the resilience of surface statistics to NULL results, their problem lies in the lack of flexibility. Through our qualitative analysis of LLMs-as-judges, we have pinpointed the underlying causes of the prevalent mismatch between BLEU and COMET. High COMET and low BLEU scores for particular segments are mostly influenced by lexical variations, word-order variations, and morphological variations. All these variations ensure semantic accuracy; however, they heavily penalise BLEU, which focuses on a perfect match. Utilising LLMs as judges (Claude, in particular, since it appeared to have better classification than Gemini) helps us determine the difference between legitimate variations and translation errors.

7 Conclusion & Future Work

We apply the QRA apparatus to two editions of WMT Low-Resource Indic MT task, treating six metrics and two LLM judges as measuring systems rather than candidates for correlation with a human gold standard that does not exist for languages. Three findings hold. First, script coverage govern neural metrics behaviour, that is, moving from Manipuri Bengali script to Meitei Mayek shifts BERTScore and COMET score by 20 points, with language and context fixed and no metric preserves system ranking across the change, though the effect reverses in the opposite direction. Second, embedding metrics compress, whereas BERTScore separates systems least and retains 0.81 of the top system's score on degenerate output, whereas BLEU and METEOR sit near the floor. Third, QRA's Type I measure does not transfer across instruments; the rank-based alternative we propose answers a related question at the cost of magnitude information.

Limitations

We report no correlation with human judgement, because none exists at scale for these languages. Our validity claims instead rest on conditions where the correct answer is known *a priori*, invariance under a change of script, and

near-zero scores for degenerate output, which show that certain instruments behave implausibly, but do not rank the instruments by fidelity to human preference. LLM-as-a-judge helps qualitatively identify whether metric discrepancies stem from word choice, morphology, or word order. They also provide natural-language explanations that make these differences easy to interpret. Finally, the released table gives system-level scores only, so our unit of analysis is the published ranking rather than the individual translation.

Ethics Statement

We identify submissions with anomalous score patterns and report them as observations about metric behaviour, not participant conduct, making no claim about intent; affected teams are anonymised, and we notified the task organisers before submission. The languages studied are low-resource technically and minoritised socially, and our central finding, that neural metrics inflate scores for text in scripts outside their pretraining coverage, means published leaderboards may overstate the state of the art for precisely the languages least able to detect the error. We frame our recommendations toward more conservative reporting rather than toward discouraging work on these pairs, and caution against adopting LLM judges as an evaluation standard for languages whose coverage in the judge model is undocumented. We use only publicly released data in accordance with its terms, and release our digitised table, code, prompts and raw judge outputs for verification. No human subjects were involved.

References

Anthropic. 2024. Introducing the next generation of claude. *Anthropic Blog.*

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anya Belz. 2022. A metrological perspective on reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz. 2025. Qra++: Quantified reproducibility assessment for common types of results in natural language processing.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2026. Quantified reproducibility assessment for four common types of evaluation results in nlp/ml. *Computational Linguistics*, pages 1–13.

Anya Belz, Craig Thomson, and Javier González Corbelle. 2026. The shared task on reproducibility of evaluations in NLP (ReproNLP) 2026: Overview and results. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 1055–1070, San Diego, California, USA. 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*, pages 8440–8451, Online. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages

733–774, Online. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

David E. Giles. 2022. Unbiased estimation of the standard deviation for non-normal populations. *WSEAS Transactions on Mathematics*, 21:119–121.

Wayne H. Holtzman. 1950. The unbiased estimate of the population variance and standard deviation. *The American journal of psychology*, 63 4:615–7.

Rudali Huidrom, Vikas Kumar, Hoomexsun Pangsatabam, Kabi Konthoujam, Nongmaithem Zeno, Nicholas Khumukcham, Mangalton Okram, Justice Konjengbam, Sai Varun Jamalpoor, Alex D Nelson Khangembam, Anand Konjengbam, and Vikram Goyal. 2026. PANINI: Improving low-resource english ↔ manipuri translation in two scripts via rslora and three-stream self-augmentation. In *Proceedings of the Eleventh Conference on Machine Translation (WMT 2026)*. Association for Computational Linguistics.

Joint Committee for Guides in Metrology. 2012. International vocabulary of metrology – basic and general concepts and associated terms (VIM).

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource Indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 532–553, Suzhou, China. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 2026. Findings of WMT 2026 shared task on low-resource Indic languages translation. In *Proceedings of the Eleventh Conference on Machine Translation (WMT26)*, Budapest, Hungary. 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sarvam AI. 2025. Sarvam models: Speech, text & translation ai. `https://www.sarvam.ai/blogs/sarvam-translate`.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No

language left behind: Scaling human-centered machine translation. *arXiv preprint*.

E. J. Williams. 1959. The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):396–399.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Method

A.1 QRA as a measurement apparatus

Meta-evaluation of MT metrics is conventionally posed as correlation with human judgment. That formulation presupposes a gold standard, and for the languages in this task, none exists at scale. We therefore adopt a measurement-theoretic apparatus in its place. The Quantified Reproducibility Assessment (QRA) framework (Belz et al., 2022; Belz, 2022), in the form given by Belz and Thomson (2026), takes its definitions from metrology rather than from NLP practice and, in doing so supplies something the correlation paradigm does not: a taxonomy of result types, each with a matching continuous-valued measure, reported at three fixed levels of granularity. It is this apparatus, rather than the reproducibility application for which it was designed, that we employ.

Grounding. QRA rests on the International Vocabulary of Metrology (Joint Committee for Guides in Metrology, 2012), under which reproducibility is the precision of a set of measured quantity values obtained under reproducibility conditions, and §2.24 of the VIM lists *different measuring systems* among those conditions. This is the clause on which our adaptation turns. We hold the measurand fixed, translation quality, and hold the objects fixed, the submitted systems, or in the adjudication study the individual segments, while varying the measuring system across the six automatic metrics and, in Experiment 5, the two judge models. What QRA then quantifies is the precision of a set of measurements taken by different instruments of the same quantity, which is a well-formed metrological question and a different one from the reproducibility question the framework was built to answer. We state the substitution explicitly here because it governs what the results in §5 can and cannot support: QRA measures the *agreement* of instruments, and agreement is not correctness. Experiments 3 to 5 supply the validity claims separately.

Result types. QRA distinguishes four types of evaluation results. **Type I** results are single numerical scores, such as a BLEU score or a mean quality rating. **Type II** results are sets of comparable numerical scores, such as the scores of all systems on one language pair. **Type III** results are categorical labels attached to text spans, such as marked-up translation errors. **Type IV** results are findings about differences between systems, such as the claim that one system outranks another. The distinction matters because relative claims can be stable where absolute ones are not, and because the four types require different measures.

Measures. For Type I results, QRA uses the bias-corrected coefficient of variation (CV*), chosen because it is scale-free and therefore comparable across studies. Given measured values $v_1, \ldots, v_n$ with sample mean m and unbiased standard deviation s^*,

$$\mathrm{CV}^* \;=\; 100 \cdot \frac{s^*}{|m|}, \qquad s^* = c_n s, \tag{1}$$

where the correction factor c_n of Holtzman (1950) removes the bias remaining in s after Bessel's correction,

$$c_n \;=\; \sqrt{\frac{n-1}{2}}\,\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \tag{2}$$

Where normality is doubtful, Belz and Thomson (2026) provide the non-parametric alternative c_n^* of Giles (2022),

$$c_n^* \;=\; \frac{8n(n-1)}{8n(n-1) - (n-1)(\kappa-3) - 2n}, \tag{3}$$

with κ the sample kurtosis coefficient.

For Type II results QRA uses Pearson's r, Spearman's ρ and Kendall's τ, generalising to Kendall's W where more than two sets of scores are compared. We report all three of r, ρ and τ_b rather than selecting one, following the framework's rationale that each affords a different perspective on the same scores. Type II measures do not apply at the system level, where only a single score exists, nor at the study level, where score sets lie on incommensurable scales.

For Type III results, QRA uses Cohen's κ for two label sets and Krippendorff's α for more, treating all labels from one measurement as issuing from a single rater. Automatic metrics emit no span-level labels, so Type III is reported as n/a for the metric analysis, a situation shared by all three worked examples in Belz and Thomson (2026), and is computed only over the judge error categories in Experiment 5.

For Type IV results QRA defines P, the proportion of identical pairwise system ranks across all pairs of measurements,

$$P = \frac{1}{|\mathcal{E}||\mathcal{S}|} \sum_{(E_i,E_j)\in\mathcal{E}} \sum_{(S_m,S_n)\in\mathcal{S}} \mathbb{1}\left[\sigma_{ij}^{mn}\right], \quad (4)$$

where σ_{ij}^{mn} denotes the condition $\operatorname{sgn}(M_{E_i}(S_m) - M_{E_i}(S_n)) = \operatorname{sgn}\big(M_{E_j}(S_m) - M_{E_j}(S_n)\big)$, $\mathcal{E}$ is the set of distinct measurement pairs, $\mathcal{S}$ the set of distinct system pairs, and $M_E(S)$ the value returned by measurement E for system S. P is the measure that speaks most directly to shared-task practice, because it asks how often two instruments would produce the same published finding.

Levels. Every measure is reported at three levels: the *system* level, comparing values obtained for one system by different instruments; the *quality criterion* level, which in our instantiation is the language pair; and the *study* level, aggregating across pairs. Types II and IV do not apply at the system level, and Type II does not apply at the study level. We follow the QRA3 reporting template so that our tables are directly comparable with the assessments reported elsewhere.

Instantiation. Objects are the submitted systems; the measurand is translation quality; measuring systems are BLEU, METEOR, TER, chrF++, BERTScore and COMET, joined in Experiment 5 by two judge models; quality criteria are language pairs. As QRA requires, all scales are oriented so that higher is better and shifted to start at zero: TER is re-expressed as $100 - \text{TER}$, and we note where its unboundedness above leaves the resulting scale only nominally bounded below.

A.2 Adapting Type I to cross-instrument comparison

Types II and IV transfer to cross-instrument use without modification, being defined over rankings and over sets of comparable scores, both of which are well formed when the measurements come from different instruments. Type I does not, and characterising the failure is a contribution of this paper.

The failure. CV* is a ratio of dispersion to central tendency, and is therefore interpretable only when the values entering it lie on a common scale with a common zero. This holds by construction in the reproducibility setting, where the same instrument is applied repeatedly. It does not hold across instruments. On our data BERTScore occupies an interquartile range of [81.9, 94.5] while BLEU occupies [7.4, 22.4], so the dispersion of a single system's six scores is dominated by the offsets between instrument scales rather than by any disagreement about that system's quality. The resulting study-level CV* of 100.4 sits far outside the framework's poor band for metrics (CV$^* > 5$), but the number reflects only the fact that the instruments are differently calibrated.

The obvious remedy fails for a second and less obvious reason. Min–max normalisation of each instrument within a language pair maps the lowest-ranked system to zero on every instrument on which it is last. Because instruments largely agree at the bottom of the table, the same system is frequently last under most or all of them: across the 148 primary submissions in pairs with $n \geq 5$, 13 systems score zero on at least five of the six instruments, and five score zero on all six. For those five the normalised mean is exactly zero, so the denominator of Eq. 1 vanishes and CV* is undefined rather than merely large; for systems

last on five of six it reaches 257.4. Nineteen systems exceed $\mathrm{CV}^* = 100$ against a median of 14.1, so the aggregate is governed by a tail of floor effects rather than by any property of the instruments. Any normalisation that anchors the empirical floor at zero inherits the pathology.

A rank-based Type I. The measure retains the interpretation Type I is meant to carry: it asks how much the instruments disagree about where a given system stands. Let $\mathcal{I}$ be the instrument set with $|\mathcal{I}| = k$, let $\mathcal{S}_q$ be the systems submitted to language pair q with $|\mathcal{S}_q| = n_q$, and let $r_i(s)$ denote the rank of system s under instrument i, computed with ties averaged and 1 denoting the best system. The system-level measure is

$$\mathrm{CV}^*{}_{\mathrm{rank}}(s) \;=\; 100 \cdot \frac{c_k \cdot \mathrm{sd}\big(\{r_i(s)\}_{i\in\mathcal{I}}\big)}{\mathrm{mean}\big(\{r_i(s)\}_{i\in\mathcal{I}}\big)}, \quad (5)$$

aggregated to the quality-criterion level as the mean over $s \in \mathcal{S}_q$ and to study the level as the mean over pairs, following the framework's own aggregation convention for Type I.

Ranks share a scale across instruments by construction, which is the property of the shift-to-zero requirement exists to secure, and the mean rank is bounded below by 1, so the denominator cannot collapse. The measure retains the interpretation Type I is meant to carry: it asks how much the instruments disagree about where a given system stands.

Remark on scale invariance. CV^* is invariant to multiplication of its argument by a positive constant. Rescaling ranks to quantile positions $r_i(s)/(n_q{+}1)$, which one might adopt to make values comparable across pairs of unequal size, therefore leaves Eq. 5 unchanged. Comparability across pairs cannot be recovered by normalisation and must be obtained by design.

Size control. Rank dispersion is bounded by the number of systems, so $\mathrm{CV}^*{}_{\mathrm{rank}}$ is not directly comparable between a pair with fifteen submissions and one with five. We therefore report, in addition to the uncontrolled measure, a size-controlled variant in which each language pair is subsampled without replacement to a fixed K systems, ranks are recomputed within the subsample, and the measure is averaged over B draws. Comparing the two versions is itself informative: where the controlled and uncontrolled orderings coincide, the association between instrument disagreement and entrant count is substantive rather than an artefact of the measure's bounds. We report both, together with the correlation between quality-criterion-level $\mathrm{CV}^*{}_{\mathrm{rank}}$ and n_q.

A.3 Statistical inference protocol

QRA specifies measures, but not inference, and the measures are point estimates computed on small samples. We therefore fix the following protocol in advance and apply it uniformly across Experiments 1 to 5.

Interval estimation. All reported values of CV^*, r, ρ, τ_b, W and P carry bias-corrected and accelerated bootstrap intervals with $B = 1000$ resamples. Resampling is over systems for system-level analyses and over segments where segment-level data exists. Where an experiment compares instruments, resampling is paired: the same resampled units are used for every instrument, so that intervals reflect uncertainty about the population rather than independent noise in each estimate.

Comparing dependent correlations. Several of our questions take the form: does instrument X agree with a given reference judgment significantly better than instrument Y? The two correlations share a variable and are therefore dependent, and comparing them by inspecting overlapping intervals is invalid. We use the Williams test for the difference between dependent correlations sharing one variable (Williams, 1959).

Comparing P values. Differences in P between instrument pairs are assessed by a paired permutation test over system pairs with 10,000 permutations, since the underlying units, pairwise rank agreements, are not independent across instruments.

Ties. Eq. 4 scores a pair as agreeing when both measurements return $\mathrm{sgn} = 0$, so instruments that produce many ties accrue agreement by construction. This is immaterial among the automatic metrics, which almost never tie at four significant figures, but it is material for the judge scores, which are elicited on five-point scales. Left unaddressed it would

make the coarser instrument appear the more concordant. We therefore adopt the following rule and apply it throughout: a tie under both measurements counts as agreement, a tie under one and a difference under the other counts as disagreement. We use τ_b rather than τ_a for the same reason, and report the tie rate alongside every P so that readers can assess the extent to which a given value is tie-driven.

Multiple comparisons. Each experiment generates a family of tests: one per instrument pair, or one per language pair. We apply the Holm–Bonferroni correction within each family and report both raw and adjusted p-values.

Statistical power. With three to fifteen systems per language pair, system-level correlation estimates are underpowered and their intervals are correspondingly wide. We report intervals throughout and do not interpret differences between instruments whose intervals overlap, including where a point estimate would suggest an ordering. Language pairs with fewer than five primary submissions are reported individually and excluded from aggregates. Where segment-level data is available, in Experiment 5, it carries the inferential weight; elsewhere our claims rest on the consistency of an effect across many language pairs rather than on its significance within any one.

Preregistration of analysis choices. Coverage-tier assignments (§4.3), the divergence-sampling rule and random seed (§4.6), the degeneracy threshold of BLEU $\leq$ 1.0 (§4.4), and the values of K and B in the size-controlled Type I are fixed before analysis and released with our code and digitised data, so that none can be adjusted to the observed effect.

B Linguistic background for languages

The languages in the shared task are highly diverse in terms of typology, morphology, and orthography among Northeast Indian languages. These languages belong to three different language families, which posed challenges due to differences in their structural properties.

1. **Tibeto-Burman family**: The languages from these family are Manipuri, Bodo, Mizo, Kokborok, Karbi, and Tagin. These languages have inflectional and derivational morphology where the case, tense, aspect, and modality are marked through a large number of suffixes. Mostly, the syntax follows the Subject-Object-Verb order.

2. **Indo-Aryan family**: Languages from this family include Assamese and Nagamese (creole). While Assamese is one of the complex Indo-Aryan languages with inflectional morphology, Nagamese is an Indo-Aryan-based creole and lingua franca of Nagaland, known for its contact-induced simplification.

3. **Austro-Asiatic family**: Language from this family is Khasi, which, unlike the neighbouring SOV Tibeto-Burman languages, has SVO typology and analytic morphosyntax with prepositional case marking.

C LLM-as-a-Judge Qualitative Examples

As an example of how segment-wise diagnosis is possible using our proposed model of LLM as a Judge, we include below some qualitative examples chosen from 100 cases that had very high COMET scores but relatively low BLEU scores. Here, the evaluators' task was to classify each divergence into one of the five predefined categories: Lexical Variation, Word Order Variation, Morphological Variation, Named Entity Variation, or Genuine Meaning Error.

Table 1 and Table 2 detail these classifications for the English-to-Manipuri (Bengali script) and Manipuri-to-English translation directions, respectively.

D Additional Results (WMT2026)

Figures below use the WMT2026 results (20 language pairs, 6 metrics: BLEU, METEOR, TER, chrF++, BERTScore, COMET; primary submissions only).

Figures below use the WMT2026 results (20 language pairs, 6 metrics: BLEU, METEOR, TER, chrF++, BERTScore, COMET; primary submissions only).

Table 1: Examples of Manipuri (mni)→English (en) translations taken from WMT 2025. Error types and explanations were generated by an LLM and verified by a human.

Source (mni)	Reference (en)	Translation (en)	Error type	Explanation
লমকোয়বসিংনা হায়দ্রবাদকি চৌমহল্লা পেলেসতা খোংনা চল্লি অমসুং নিজামগি মতমগি ইলিজেন্স অমসুং পুবারি অদু উবা ফংই।	Tourists stroll through Hyderabad's Chowmahalla Palace witnessing Nizam-era elegance and history.	Tourists walk through the Chowmahalla Palace in Hyderabad and witness the elegance and grandeur of the Nizam era.	Lexical Variation	The MT uses 'walk' instead of 'stroll' and 'grandeur' instead of 'history' to translate the source terms, which are acceptable lexical variations that maintain the original meaning.
চহি কয়াগি মতুংদা মিয়াম্না পাম্নবা সিতকম্ অসি অনৌবা সিজনগিদমক্ হল্লক্লে।	Popular sitcom returns for new season after several years' hiatus.	After many years, the popular sitcom is back for a new season.	Word Order Variation	The MT and reference convey the same meaning, but the MT uses a different word order by placing the temporal phrase at the beginning of the sentence.
গুকেসকি মপা রজনিকান্তনা ভারতকি সেনিয়র গ্রেন্দমাস্তর অমদা ক্রেদিৎ পিখি।	Gukesh's father Rajnikanth gave credit to a senior Indian Grandmaster	Gukesh's father Rajnikant gives credit to one of India's senior grandmasters.	Morphological Variation	The MT renders the past-tense পিখি ('gave') as present 'gives', a tense/inflection mismatch; other differences (Rajnikant spelling, 'one of India's senior grandmasters') are trivial surface variants preserving meaning.
ব্রয়ান্ জোনসন্না মহাক্কি প্রমান্ লৈরবা ত্রাভেল্ রিকোভরি ত্রিকসিং সেয়র তৌরি।	Bryan Johnson Shares His Proven Travel Recovery Tricks	Brian Johnson shares his proven travel recovery tricks.	Named Entity Variation	The only substantive difference is the transliteration of the first name: reference 'Bryan' vs MT 'Brian'; the rest differs only in sentence-case and final period.
দিপিকা পাদুকোন্না নহা ওইরিবা মিওইসিংদা মতেং পাংনবা মেন্তেল্ হেল্থ ইনিসিয়েতিব হৌদোকখ্রে।	Deepika Padukone launches mental health initiative to support young individuals.	Deepika Padukone has launched the Mental Health Initiative to help young women.	Genuine Meaning	The source মিওইসিং means 'people/individuals' generally, but the MT narrows it to 'young women', introducing a gender that is not in the source; the rest matches the reference.

D.1 Experiment 1: Discriminative Power (CV*)

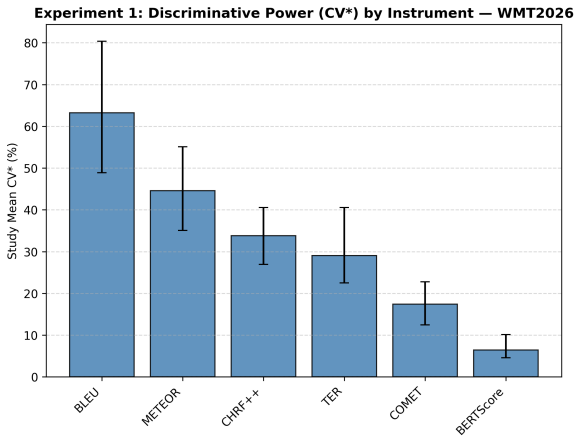


Figure 7: Study-level CV* per instrument across all WMT2026 pairs (95% CI). Higher = more discriminative.

In the fig 7 surface metrics (BLEU, METEOR) show the highest CV*; neural metrics (COMET, BERTScore) the lowest, so they discriminate between systems less. This ordering holds across both submission categories.

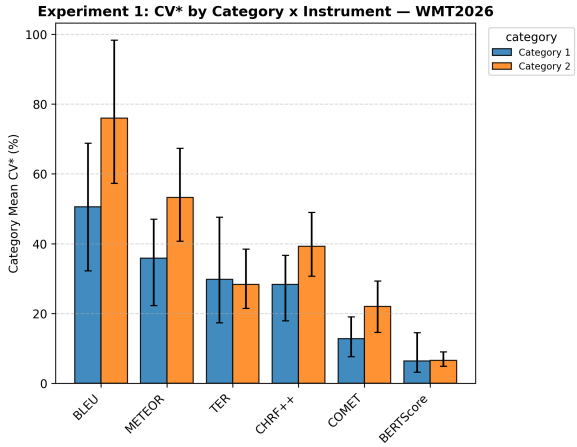


Figure 8: CV* by submission category (Category 1 vs. Category 2), per instrument.

D.2 Experiment 2: Consensus Centrality

In fig 9, 10, and table 4 metric agreement is high ($W > 0.8$) for most pairs but drops for Karbi, Tagin, and Kokborok. chrF++ and BERTScore are the most central metrics; TER agrees least with the rest.

Table 2: Examples of English (en)→Manipuri (mni) translations taken from WMT 2025. Error types and explanations were generated by an LLM and verified by a human.

Source (en)	Reference (mni)	Translation (mni)	Error type	Explanation
Tourists visit Ajmer's Dargah seeking blessings at Rajasthan's famous Sufi pilgrimage site.	লমকোয়বসিংনা রাজস্থানগি মমিং লৈরবা সুফি তিরথ মফম্ অসিদা থৌজাল্ লৌদুনা অজমেরগি দারগাহদা চল্লি।	লমকোয়বশিংনা রাজস্থানগী মমিং চৎলবা সুফি লাই খুরুম্ফমদা থৌজাল ফংনবগীদমক অজমেরগী দর্গাদা লাকই।	Lexical Variation	The MT conveys the same proposition, differing only in word choice: 'মমিং চৎলবা' vs 'মমিং লৈরবা' for famous, 'লাই খুরুম্ফম' vs 'তিরথ মফম' for pilgrimage site, and 'লাকই' (come) vs 'চল্লি' (go), plus a spelling variant of Dargah.
Lok Sabha telecommunication policies debates to address 5G rollout challenges.	লোক্ সভা তেলিকম্যুনিকেসন্ পোলিসিসিংগি মতাংদা ৫জি রোল্ আউৎ চেলেন্জসিং কোকহন্নবা দিবেৎ তৌরি।	5জি রোলআউট চেলেঞ্জশিং থেংননবা লোক সভা তেলিকম্যুনিকেসন পোলিসীশিংগী দিবেতশিং	Word Order Variation	The MT uses a different word order by placing the purpose clause at the beginning and uses a noun phrase structure instead of a full sentence, but the meaning remains identical.
Chelsea's 11 against Newcastle was the youngest in the league.	চেলসিনা ন্যুক্যাসলগা থেংনখিবা ১১ অদু লিগ অসিগি খ্বাইদগি নৌবা ওইখি।	ন্যুকেসলগী মায়োক্তা চেল্সিয়াগী 11 অসি লিগ অসিদা খ্বাইদগী নহা ওইরিবনি।	Morphological Variation	Same proposition; the clearest difference is the verb inflection (reference past ওইখি vs MT present/continuative ওইরিবনি), plus minor variants in the transliteration of Chelsea/Newcastle, the numeral script (১১ vs 11), and নৌবা vs নহা for 'young'. Competing label would be lexical_variation given those word-choice/spelling differences.
Pilgrims meditate at Bodh Gaya's temple discovering peace under Buddha's sacred tree.	তিরথ চৎপসিংনা বুদ্ধগি সেংলবা উগি মখাদা সান্তি থিদুনা বোধ গয়াগি মন্দিরদা ধ্যান্ তৌই।	তীর্থযাত্রীশিংনা বুদ্ধ গয়াগী লাইশংদা ধ্যান তৌদুনা বুদ্ধগী শেংলবা উগী মখাদা শান্তি ফংবা ঙম্লে।	Named Entity Variation	The MT renders "Bodh Gaya" as "বুদ্ধ গয়" instead of the reference's "বোধ গয়", a transliteration variant of the same place; other differences (তীর্থযাত্রীশিং vs তিরথ চৎপসিং, "শান্তি ফংবা ঙম্লে" for "discovering peace") are acceptable lexical/aspectual variation.
Wildlife Sanctuaries in Karnataka: are the greatest Haven for Nature Lovers	করনাতকাদা লৈবা বাইল্দলাইফ সেন্ত্রেলসিং: মহৌসাবু নুংসিজবসিংগি খ্বাইদগি চাউবা হেভননি।	কর্নাতকাগী ৱাইল্ডলাইফ সেঙ্কচুয়ারিশিং: মহৌশাগী ফজবশিংগী খ্বাইদগী চাওবা মফম ওইরি	Genuine Meaning	The MT renders "nature lovers" as "মহৌশাগী ফজবশিং" (nature's beautiful things/beauties) instead of people who love nature (reference: নুংসিজবসিং), so the beneficiary of the 'haven' is misrepresented; 'haven' is also flattened to the generic 'মফম' (place).

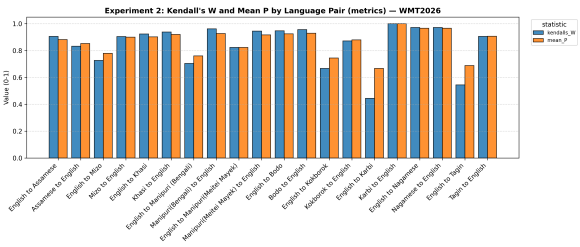


Figure 9: Kendall's W and mean pairwise agreement P per language pair (metrics only).

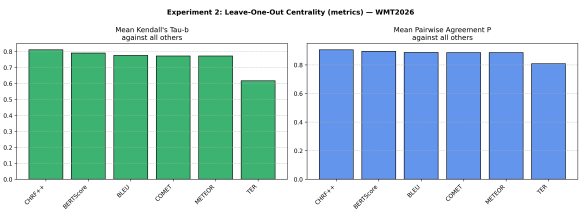


Figure 10: Leave-one-out centrality: mean Tau-b and P of each instrument against all others.

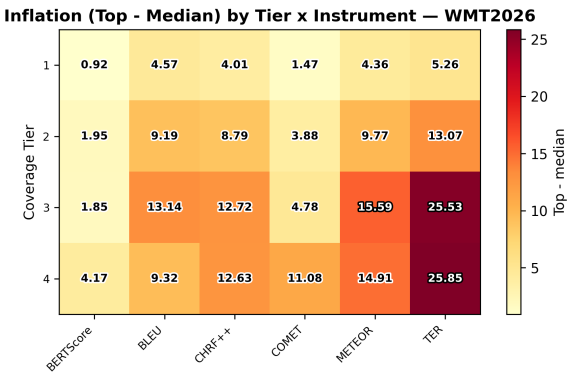


Figure 11: Tier × instrument heatmap.

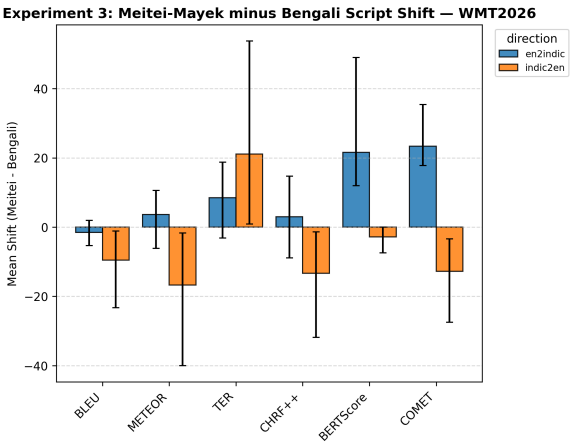


Figure 12: Paired Meitei-Mayek minus Bengali score shift, per instrument and direction (95% CI).

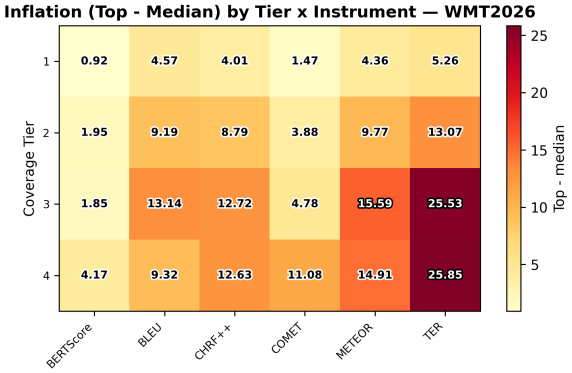


Figure 13: Tier × instrument heatmap.

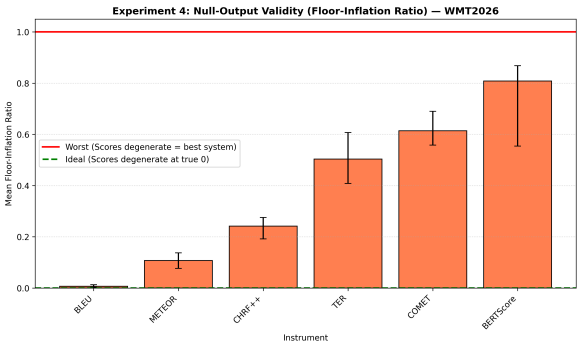


Figure 14: Mean floor-inflation ratio per instrument (0 = ideal, 1 = worst), 95% CI.

Table 3: Study-level CV* (%) and mean observed range by instrument, WMT2025/WMT2026 (bootstrap 95% BCa CI, B=1000).

WMT2026			WMT2025		
Instr.	CV*	Range	Instr.	CV*	Range
BLEU	63.3	18.7	BLEU	73.9	9.4
METEOR	44.6	34.6	ROUGE-L	63.4	0.2
chrF++	33.8	30.8	METEOR	63.1	0.3
TER	29.1	68.5	Cos Sim.	40.7	0.4
COMET	17.4	25.5	chrF	38.1	24.2
BERTScore	6.5	14.1	TER	36.0	117.6

Table 4: Leave-one-out centrality (metrics scope): mean τ_b and P vs. all other instruments, study level.

WMT2026			WMT2025		
Instr.	τ_b	P	Instr.	τ_b	P
chrF++	0.811	0.905	Cos Sim.	0.892	0.946
BERTScore	0.790	0.894	chrF	0.886	0.942
BLEU	0.775	0.887	METEOR	0.883	0.941
COMET	0.773	0.886	BLEU	0.875	0.937
METEOR	0.772	0.885	ROUGE-L	0.786	0.890
TER	0.617	0.808	TER	0.704	0.851

D.3 Experiment 3: Script Sensitivity & Coverage-Tier Inflation

In the table 5 BERTScore and COMET inflate Meitei-Mayek scores on English→Indic, but the effect disappears on Indic→English. In the fig 12, 11, and 13 Surface metrics inflate steadily from Tier 1 to Tier 4; BERTScore/COMET stay comparatively flat.

D.4 Experiment 4: Null-Output Validity

As shown in Table 6 and Figure 14, BLEU and METEOR are near the ideal floor. BERTScore is the most concerning (ratio ≈ 0.81): degenerate output still scores close to the best system.

D.5 Experiment:Metrics-vs-Metrics

Table 7 summarises the Kendall's τ-b concordance between metric pairs for WMT25 and WMT26.

E Additional Results (WMT2025)

In all the experiments on WMT 2025, we use the WMT2025 results (14 language pairs, 6 metrics: BLEU, ROUGE-L, METEOR, chrF, TER, Cosine-Similarity; primary submissions only; no BERTScore/COMET, and no dual-script Manipuri split this year).

Table 5: WMT2026 paired script shift (Meitei-Mayek − Bengali), mean [95% bootstrap CI]. Bold = CI excludes 0. No WMT2025 row exists: that year has no dual-script Manipuri split.

Instrument	en→indic shift	indic→en shift
BLEU	−1.5 [−5.4, 1.9]	**−9.5 [−23.3, −1.1]**
METEOR	+3.6 [−6.2, 10.6]	**−16.8 [−40.0, −1.7]**
TER	+8.5 [−3.2, 18.7]	**+21.1 [0.9, 53.7]**
chrF++	+3.0 [−8.9, 14.7]	**−13.3 [−31.8, −1.3]**
BERTScore	**+21.6 [11.9, 49.0]**	−2.9 [−7.4, 0.0]
COMET	**+23.3 [17.8, 35.4]**	**−12.8 [−27.5, −3.4]**

Instrument	WMT26	WMT25
BLEU	0.007	0.029
METEOR	0.107	0.029
ROUGE-L	–	0.034
Cos Sim.	–	0.063 (n=1)
chrF++/chrF	0.241	0.180
TER	0.504	0.424
COMET	0.614	–
BERTScore	0.809	–

Table 6: Mean floor-inflation ratio per instrument (0=ideal, 1=worst). n = number of degenerate systems the ratio is averaged over; Cosine-Similarity (WMT25) has only n=1, so no CI could be computed (bootstrap requires $n \geq 3$). chrF++ (2026) and chrF (2025) use different pipelines and are not directly equated.

E.1 Experiment 1: Discriminative Power (CV*)

Figure 15 BLEU, ROUGE-L, and METEOR show the highest CV*; chrF, TER, and Cosine-Similarity are lower and close to each other. Confidence Intervals (CIs) are wide given the smaller (14-pair) sample.

E.2 Experiment 2: Consensus Centrality

In the figure 16 and 17, Metric agreement is very high across nearly all WMT2025 pairs ($W \geq 0.8$, with several at $W = 1.0$); Bodo→English has no metrics-only row since Cosine-Similarity is the only second metric available there. Cosine-Similarity and chrF are the most central instruments; TER agrees least with the rest.

E.3 Experiment 3: Coverage-Tier Inflation

As shown in the figure 18 and 19, TER dominates both inflation and range by a wide margin at every tier (range peaks at Tier 2, ≈286); BLEU and chrF inflate moderately; ROUGE-

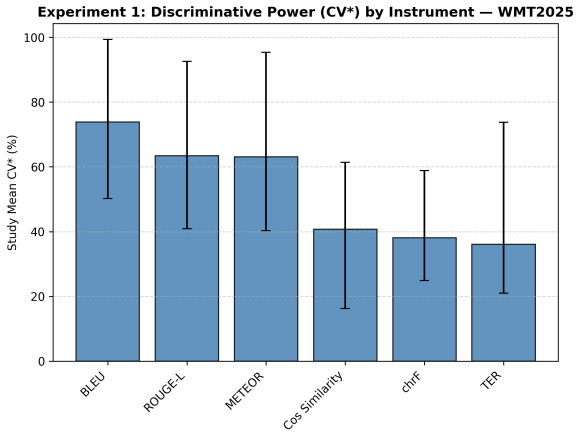


Figure 15: Study-level CV* per instrument across all WMT2025 pairs (95% CI). Higher = more discriminative.

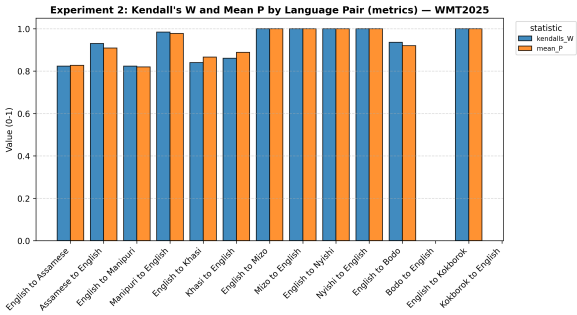


Figure 16: Kendall's W and mean pairwise agreement P per language pair (metrics only).

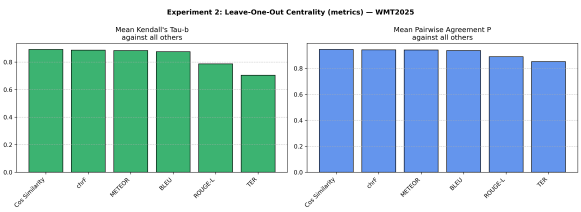


Figure 17: Leave-one-out centrality: mean Tau-b and P of each instrument against all others.

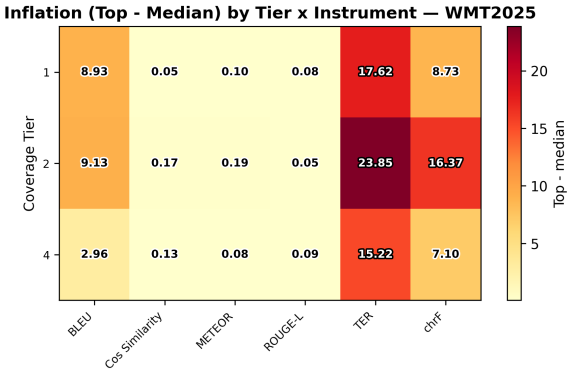


Figure 18: Inflation values from as a tier × instrument heatmap.

	WMT26	WMT25
τ-b range	0.59–0.87	0.63–0.99
Lowest pair	METEOR–TER (0.59)	ROUGE-L–TER (0.63)
Highest pair	chrF++–BERTScore (0.87)	METEOR–chrF (0.99)

Table 7: Metrics-vs-metrics Kendall's τ-b concordance summary, system level, all pairs.

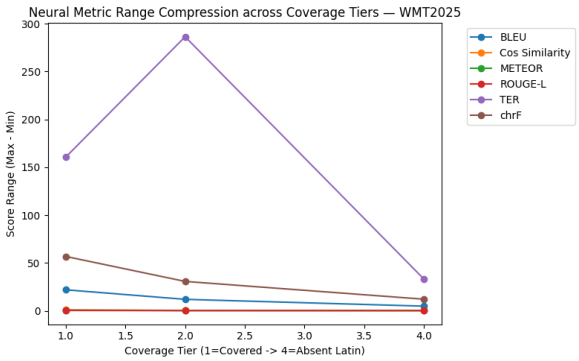


Figure 19: Raw score range (max−min) by coverage tier, per instrument.

L, METEOR, and Cosine-Similarity stay near zero throughout. No paired-script-shift figure is included here: WMT2025's Manipuri submissions are not split by script, so that comparison is unavailable for this year.

E.4 Experiment 4: Null-Output Validity

In this figure 20, BLEU, METEOR, and ROUGE-L sit near the ideal floor. TER is the most concerning (ratio ≈ 0.42, and 0.99 for English→Kokborok alone): degenerate output can score almost as high as the pair's best system.

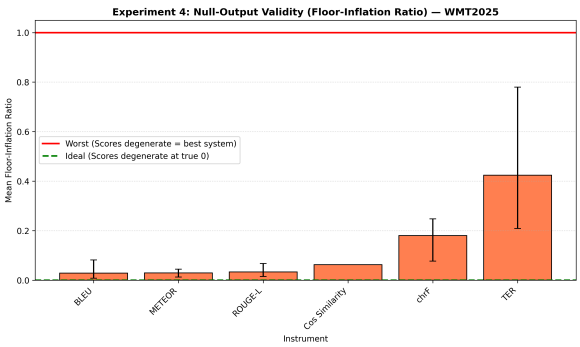


Figure 20: Mean floor-inflation ratio per instrument (0 = ideal, 1 = worst), 95% CI.

F Prompts

We use the following prompts for our *LLM-as-a-Judge* protocol to evaluate the translations. Figure 21 refers to the system prompt provided to the judge model. Here it is asked to generate adequacy, fluency and semantic equivalence scores along with a certain confidence score which shows the judge's confidence in its own judgement. It also categorizes the errors present in the translations from the given categories in fig. 22 which has 5 different categories of errors - *Lexical variation*, *Word-order variation*, *Morphological variation*, *Named entity variation* and *Genuine meaning error*. We finally include the user prompt in fig. 23 for obtaining the *LLM-as-a-Judge* output.

System Prompt

You are an expert bilingual annotator evaluating machine translation between $src_lang and $tgt_lang.
You are shown a source segment, a human reference translation, and a machine translation (MT) output. Judge the MT against the source, using the reference only as one acceptable rendering not as the single correct answer. A translation that differs from the reference in wording, word order, or morphology may still be fully correct.
Score each dimension on a 1-5 integer scale.
Adequacy — how much of the source meaning is preserved in the MT.

- 5: all meaning preserved, nothing added or lost.
- 4: nearly all meaning preserved; a minor nuance is lost.
- 3: most meaning preserved, one clear omission, addition, or distortion.
- 2: substantial meaning lost or wrong.
- 1: unrelated to the source, or meaning is inverted.

Fluency — how natural and grammatical the MT is as $tgt_lang text, judged on its own without looking at the source.

- 5: reads as though written by a native speaker.
- 4: fluent with a small awkwardness that does not impede reading.
- 3: understandable but visibly non-native; grammar or word-choice errors.
- 2: hard to read; frequent errors.
- 1: incomprehensible or not $tgt_lang.

Semantic equivalence — how close the MT is in meaning to the *reference*, ignoring surface form entirely. Two sentences that use different words, different word order, or different morphology but convey the same proposition are semantically equivalent.

- 5: identical meaning, differing only in surface form.
- 4: same meaning with a negligible difference in emphasis or specificity.
- 3: broadly the same meaning; one proposition differs.
- 2: partially overlapping meaning.
- 1: different meaning.

Taxonomy
Confidence: Also report confidence 1-5: how much you would stand behind this judgment as a whole, especially the error type. This is a claim about your own certainty, not about translation quality — a clearly bad translation you are sure about is confidence 5.

- 5: unambiguous. The difference is plain and exactly one label fits.
- 4: clear, with a small residual doubt.
- 3: defensible, but a second label fits nearly as well, or the reference itself looks questionable.
- 2: a guess between competing readings; you would not be surprised to be wrong.
- 1: you cannot properly assess this segment — the source or reference is unintelligible, truncated, or in an unexpected language.

Use the low end when you genuinely mean it. A run in which every segment is confidence 5 carries no information. In particular, drop to 3 or below when the source is ambiguous, when the reference and the MT are both defensible renderings of different readings, or when you lack the domain knowledge to tell a terminology error from a synonym.
How to decide the error type: Choose exactly one label — the one that best explains why the MT differs from the reference.

- If the MT and the reference mean the same thing, pick whichever surface category (1-4) accounts for the largest part of the difference. Do not pick "genuine meaning error" merely because the strings differ.
- Pick "genuine meaning error" only when the MT actually misrepresents the source: something is omitted, added, negated, or mistranslated such that a reader would be misinformed.
- If the MT is character-for-character identical to the reference, pick the surface category whose definition is closest to vacuous — use "lexical variation" and say in the explanation that there is no divergence.

Output: Return a single JSON object and nothing else. In `explanation`, give one or two sentences naming the concrete difference you based the error type on; if `confidence` is 3 or below, say there what the competing reading was. In `divergent_spans`, list the specific reference-vs-MT fragments that differ; leave it empty if nothing differs.
Judge only what is in front of you. Do not speculate about the system that produced the MT, and do not reward or penalise a translation for being longer or shorter than the reference.

Figure 21: System prompt used for evaluating machine translation quality using *LLM-as-a-Judge*.

Error Type

Error type (choose exactly one)

1. **Lexical variation** (`lexical_variation`) — a different but acceptable word or phrase choice: synonyms, a paraphrase, a transliteration or spelling variant, a different register, an idiom rendered literally or vice versa. Meaning is preserved.

2. **Word-order variation** (`word_order_variation`) — the same content words arranged differently: reordered clauses, active/passive alternation, fronting or extraposition, a different but valid constituent order. Meaning is preserved.

3. **Morphological variation** (`morphological_variation`) — the same lexemes carrying different inflection or affixation: tense, aspect, number, case, agreement, definiteness, honorific level, or a different derivational form of the same root. Meaning is preserved apart from the grammatical feature itself.

4. **Named entity variation** (`named_entity_variation`) — a difference confined to a proper name, number, date, measurement, currency, or other named entity: a different transliteration or spelling of a name, digits versus spelled-out numerals, a different date or unit format. The entity refers to the same thing.

5. **Genuine meaning error** (`genuine_meaning_error`) — the MT misrepresents the source: content is omitted or hallucinated, a relation is reversed, polarity or negation is wrong, or an entity refers to something different. A reader relying on the MT would be misinformed.

Figure 22: Error type taxonomy for machine translation evaluation.

Evaluation Template

Direction: $src_lang → $tgt_lang
Segment id: $seg_id
SOURCE ($src_lang):
$src
REFERENCE ($tgt_lang):
$ref
MACHINE TRANSLATION ($tgt_lang):
$mt
$scores
Evaluate the machine translation and return the JSON object.

Figure 23: User prompt for *LLM-as-a-Judge*