From Error Detection to Credit Assignment: Do xCOMET Error Spans Provide Useful Policy Credit for Chinese–Malay Translation?

Zi Chen Koo
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China
zichenkoo@gmail.com

Abstract

Sequence-level rewards assign the same group-relative advantage to every generated token, even when only a short translation span is erroneous. We ask whether xCOMET error spans provide useful token-level policy credit for Group Relative Policy Optimization (GRPO) in Chinese–Malay translation. Signed Lexical-Residual GRPO (SLR) adds a bounded, zero-mean local residual to the sequence advantage. Across three matched seeds, SLR improves reward-aligned COMET and MetricX-24, a learned metric not used in training, while chrF++ and BLEU differences remain unresolved. A blinded 300-item evaluation by the bilingual author does not resolve an adequacy or fluency preference and finds a higher omission rate under SLR (9.33% to 16.67%). A repeated span audit finds high sentence-level error detection but weak target-word localization; fewer than half of consistently marked error words receive negative residual credit. In this setting, error detection, localization, and policy-credit actionability are distinct: explainable metric spans change optimization behavior without automatically providing human-aligned token supervision.

1 Introduction

Machine translation (MT) rewards are usually assigned after a complete translation is generated. In outcome-supervised GRPO, each response receives one advantage relative to other responses sampled for the same source, and the same value is applied to every response token (Shao et al., 2024). This is efficient, but it does not say *which* token made one translation better than another. A locally wrong term can therefore be reinforced when it occurs in the best member of a weak rollout group, while correct tokens can be penalized in an otherwise below-average translation.

Explainable MT metrics expose a candidate local signal. xCOMET predicts both a sentence score and character-level error spans with severities (Guerreiro et al., 2024). Recent work maps these spans to generation tokens and directly optimizes severity rewards with REINFORCE or PPO (Ramos et al., 2026); large-scale WMT systems have also combined global MetricX feedback with token-level AutoMQM feedback (Finkelstein et al., 2025). These results establish fine-grained MT optimization. Its success still depends on whether detector boundaries identify policy actions that repair the translation.

We study this question for Chinese–Malay, an underrepresented pair with an additional target-control challenge. Malaysian Standard Malay and Indonesian share much of their lexicon and grammar, yet are distinct task targets. Generative models can answer Malay prompts in Indonesian (Nomoto, 2023), while broad-coverage language identifiers also struggle with closely related languages (Kargaran et al., 2023). Semantic adequacy and target variety must therefore be measured separately.

Our research question is: *when an explainable metric marks a translation span as erroneous, does converting that span into signed group-relative token credit improve the target construct measured by automatic and human evaluation?* Signed Lexical-Residual GRPO (SLR) adds a centered, severity-sensitive residual to the sequence advantage. The residual sums to zero within a response, is globally capped, and is disabled when spans are absent, too broad, or unmappable. This isolates signed local credit against a matched sequence-GRPO baseline.

Across three matched seeds, SLR improves MetricX-24 and COMET, while chrF++ and BLEU differences are unresolved, and it lowers the SoftLID-adjusted score. A blinded 300-item Chinese-to-Malay evaluation shows no resolved adequacy or fluency preference and a higher omission rate under SLR. In a separate 100-candidate

Proceedings of the Eleventh Conference on Machine Translation, pages 368–379
October 28-29, 2026 ©2026 Association for Computational Linguistics

unbiased span audit, xCOMET detects most sentences marked erroneous in both annotation passes but localizes their target words poorly. These results expose a gap between error detection, boundary localization, and policy-credit actionability.

This paper makes three contributions:

1. We formulate a bounded signed residual that can reverse local credit while preserving each response's mean GRPO advantage.

2. We conduct a matched three-seed evaluation on two test suites and separate reward components, MetricX, and target-language outcomes with source-clustered uncertainty.

3. We combine a 300-item blinded quality evaluation with a repeated bilingual span audit, showing that high sentence-level recall coexists with weak target-word localization and incomplete negative credit on error words marked consistently across passes.

2 Related Work

Sequence-level and group-relative MT optimization. Direct optimization of sentence-level MT metrics predates LLM reinforcement learning. Minimum-risk training estimates expected loss over sampled translations and admits non-differentiable sequence metrics (Shen et al., 2016). A subsequent REINFORCE study found that sampling strategy and the mixture of likelihood and RL objectives materially affected NMT performance, while reward shaping and a learned baseline yielded little additional gain in that setting (Wu et al., 2018).

GRPO replaces the learned value model with a baseline estimated from responses sampled for the same prompt (Shao et al., 2024). Recent MT applications generally retain a scalar sequence reward: Garcia Gilabert et al. (2025) combine terminology adherence with reference-free quality estimation; SSR-Zero uses a policy-generated quality score (Yang et al., 2026); and Su et al. (2026) combine semantic similarity with target-language consistency for Tibetan–Chinese expansion. PE-GRL instead localizes optimization through online post-editing (Shen et al., 2026). This literature establishes both sequence-level MT optimization and GRPO for MT, but does not test signed policy credit derived from post-hoc metric spans.

Outside MT, Outcome-Grounded Advantage Reshaping reallocates GRPO advantage with token outcome-sensitivity weights and renormalizes them to preserve total advantage mass (Li et al., 2026). Its nonnegative weights suppress or amplify tokens without reversing the sign of the sequence advantage. SLR instead tests signed corrections derived from external error spans.

Fine-grained MT reward. Ramos et al. (2026) align xCOMET character spans to generation tokens by overlap, assign the worst overlapping severity, and optimize the resulting local quality values with token-level REINFORCE and PPO. Their experiments cover English–German and English–French and report neural-metric and human gains, although lexical metrics can decrease in some settings. Feng et al. (2025) use approximate Monte Carlo tree search to construct token-level translation preferences and train process reward models; integrating those models into RL is left for future work. Finkelstein et al. (2025) add token-level Gemma-AutoMQM-QE advantages to a sequence-level MetricX advantage in a large proprietary multilingual system. Existing work therefore covers both direct xCOMET-severity optimization and additive global–local MT credit. We test whether a bounded, zero-mean xCOMET residual is actionable when added to group-relative sequence advantages in Chinese–Malay.

Error detection versus localization. xCOMET jointly predicts a sentence score and token-level error labels that are merged into character spans (Guerreiro et al., 2024). Its original span evaluation covers WMT language pairs including Chinese–English, but not Malay. A recent expert-span study reports substantial disagreement between human boundaries and automatic detectors: xCOMET-XXL attains approximately 0.26 character-level span F1 under that study's protocol (Thellmann et al., 2026). Its protocol and span unit differ from ours; both studies motivate separating sentence detection, boundary localization, and policy-credit actionability.

Malay and Indonesian variety control. Malay (`zsm`) and Indonesian (`ind`) are closely related standard varieties. Nomoto (2023) finds that a generative model often responds to Malay prompts in Indonesian even when it can identify the prompt language, locating the failure partly in generation rather than binary language identification. GlotLID covers many low-resource languages, but

its authors identify closely related languages and macrolanguage/variety granularity as general LID challenges (Kargaran et al., 2023). A recent comparison of OpenLID and GlotLID further shows that strong aggregate multilingual scores can hide confusion among closely related languages and that higher-precision ensembles reduce coverage (Fedorova et al., 2026). We use GlotLID as an operational control and report its decisions separately from translation quality.

3 Signed Lexical-Residual GRPO

3.1 Sequence baseline

For a source x, the policy samples a group of $G = 8$ translations $\{y_i\}_{i=1}^{G}$. The translation reward is

$$R_{\mathrm{MT},i} = 0.5C_i + 0.3F_i + 0.2B_i, \tag{1}$$

where C_i is the sentence-level xCOMET-XL score and F_i and B_i are sentence-level chrF++ and BLEU, respectively. A graded language penalty P_{lid} and script penalty P_{script} produce

$$r_i = \mathrm{clip}\left(R_{\mathrm{MT},i} - P_{\mathrm{lid},i} - P_{\mathrm{script},i}, -1, 1\right). \tag{2}$$

Within each rollout group, we compute

$$A_i = \frac{r_i - \overline{r}}{\mathrm{sd}(r) + 10^{-6}}. \tag{3}$$

Here sd is the framework's within-group sample standard deviation. Let $\mathcal{T}_i$ denote the set of active policy-token positions in response i, and let $T_i = |\mathcal{T}_i|$. The matched sequence-GRPO baseline (B4) assigns $A_{i,t}^{B4} = A_i$ for every $t \in \mathcal{T}_i$.

3.2 Why nonnegative span weighting is insufficient

A nonnegative weighting rule $A_{i,t} = A_i w_{i,t}$ can redistribute update magnitude, but $\mathrm{sign}(A_{i,t}) = \mathrm{sign}(A_i)$ whenever $w_{i,t} > 0$. Consequently, a known local error cannot be penalized when its response has $A_i > 0$. Conversely, correct words in a response with $A_i < 0$ cannot be protected by a sign-locked weight. SLR introduces a signed residual specifically to test whether relaxing this constraint is useful.

3.3 Lexical quality residual

We segment a Malay response by whitespace and punctuation and a Chinese response with Jieba. After excluding punctuation and units that overlap no active policy token, let $\mathcal{L}_i$ denote the retained lexical units and let $L_i = |\mathcal{L}_i|$. Every retained unit ℓ starts with quality $q_{i,\ell} = 8$. If an xCOMET span overlaps the unit, we assign the worst overlapping severity using

$$q(\mathrm{correct}, \mathrm{minor}, \mathrm{major}, \mathrm{critical}) = (8, 4, 2, 1), \tag{4}$$

following the severity mapping of Ramos et al. (2026). We define the mean over retained units as $\overline{q}_i = L_i^{-1} \sum_{\ell \in \mathcal{L}_i} q_{i,\ell}$ and use the fixed severity range:

$$u_{i,\ell} = \frac{q_{i,\ell} - \overline{q}_i}{7}. \tag{5}$$

Unlike per-response RMS normalization, this retains the difference between minor, major, and critical errors in the common case where a response contains only one error severity.

For each $\ell \in \mathcal{L}_i$, let $\mathcal{T}_{i,\ell} \subseteq \mathcal{T}_i$ be its overlapping active policy tokens and let $K_{i,\ell} = |\mathcal{T}_{i,\ell}|$. The raw token residual is

$$g_{i,t}^{\mathrm{raw}} = \sum_{\substack{\ell \in \mathcal{L}_i: \\ t \in \mathcal{T}_{i,\ell}}} \frac{T_i}{L_i} \frac{u_{i,\ell}}{K_{i,\ell}}, \qquad t \in \mathcal{T}_i, \tag{6}$$

with $g_{i,t}^{\mathrm{raw}} = 0$ outside $\mathcal{T}_i$. Dividing by $K_{i,\ell}$ makes each lexical unit's total contribution invariant to subword fragmentation. Let $M_i = \max_{t \in \mathcal{T}_i} |g_{i,t}^{\mathrm{raw}}|$. One response-level scale,

$$\kappa_i = \begin{cases} 1, & M_i \le 2, \\ 2/M_i, & M_i > 2, \end{cases} \qquad g_{i,t} = \kappa_i g_{i,t}^{\mathrm{raw}}, \tag{7}$$

caps $|g_{i,t}| \le 2$ without changing severity ordering or the zero-sum property. The SLR token advantage is

$$A_{i,t}^{SLR} = A_i + \lambda g_{i,t}, \qquad \lambda = 0.25. \tag{8}$$

Therefore $|\lambda g_{i,t}| \le 0.5$ and

$$\sum_{t \in \mathcal{T}_i} g_{i,t} = 0, \qquad \frac{1}{T_i} \sum_{t \in \mathcal{T}_i} A_{i,t}^{SLR} = A_i. \tag{9}$$

At $\lambda = 0$, SLR is exactly B4. For $\lambda > 0$, zero mean preserves the response-average advantage, while the token-level policy gradient changes according to the localized residual.

3.4 Alignment and reliability gates

The live mapper reconstructs the exact generated byte stream from policy token IDs and maps character spans through UTF-8 byte overlap. This handles Chinese characters split across multiple

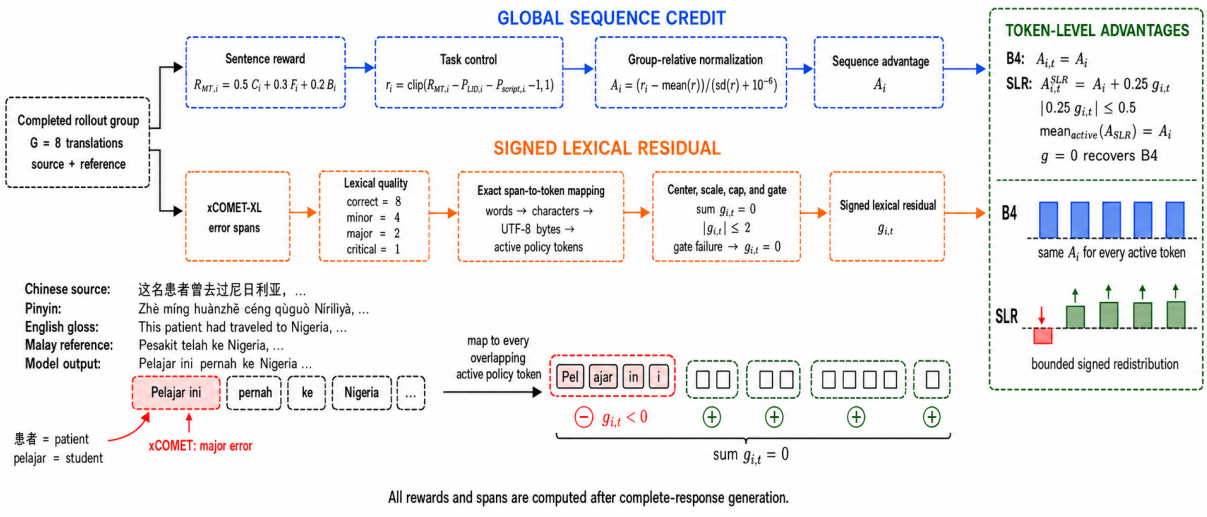

Figure 1: Training and credit assignment after complete-response generation. The upper branch produces sequence advantage A_i; the lower maps xCOMET spans to bounded, zero-mean residual $g_{i,t}$. B4 broadcasts A_i, whereas SLR adds $0.25g_{i,t}$. In the frozen Chinese–Malay example, xCOMET marks "Pelajar ini" ("this student") as a major mistranslation of "this patient," and the mapper assigns its residual to every overlapping active token. Token pieces and bar heights are schematic; the right panel shows a possible sign reversal, but a negative residual need not make $A^{SLR}_{i,t}$ negative.

byte-level BPE tokens and non-canonical token sequences without re-tokenizing the displayed response. If a model emits an invalid UTF-8 sequence, the sample is counted and receives $g = 0$; any other response/token mismatch raises an error.

We also set $g = 0$ when there are no spans, no span overlaps an active lexical unit, all mapped units have the same quality, more than 60% of eligible words are marked erroneous, or the wrong-script penalty exceeds 0.30. Thus the intervention is local only when the detector output is both non-trivial and placeable on policy tokens.

4 Experimental Setup

4.1 Training data

The RL corpus contains 4,691 Chinese–Malay translation instances: 2,349 target Malay and 2,342 target Chinese. Malay sentences were collected from publicly accessible news, government, and general web pages. Source filtering removed malformed and machine-translated web text. Selected sentences were translated into Chinese with multiple LLMs, including few-shot GLM-4.7, and the resulting pairs were cleaned for completeness and bidirectional use.

Against the final corpus, neither evaluation suite contains an exact normalized sentence match. Shared 8-grams occur in 4/1,012 FLORES+ pairs and 6/1,997 NTREX pairs, all on the Chinese side.

4.2 Evaluation data

We use the public FLORES-200/FLORES+ `devtest` split for Malaysian Standard Malay (`zsm_Latn`) and Simplified Chinese (`zho_Hans`), containing 1,012 parallel items (Goyal et al., 2022; NLLB Team, 2024). FLORES-200 is released under CC BY-SA 4.0.

NTREX-128 contributes 1,997 official Malay (`msa`) and Chinese (`zho-CN`) news pairs (Federmann et al., 2022). NTREX references were translated independently from English; using them reciprocally for Chinese–Malay evaluation introduces translationese and occasional reference errors.

Both suites informed reward development. They evaluate the matched B4/SLR comparison under the same fixed reward but are not fully held out from the overall reward-design process.

For both suites, we translate every item in both directions with Chinese and Malay instruction prompts. This yields four conditions per item: 4,048 FLORES+ rows and 7,988 NTREX rows per model and seed.

4.3 Model and optimization

Both conditions initialize from the public `zai-org/GLM-4-9B-0414` checkpoint at revision `645b848`, released under the MIT license (Team GLM, 2024).

Each run uses four A100 GPUs with tensor parallelism 2. We train for 420 rollout steps, sampling 12 prompts and eight responses per prompt (96 responses per step). The maximum prompt and response lengths are 4,096 and 768 tokens, with sampling temperature 1.0. Adam uses learning rate 2×10^{-6}, cosine decay to 2×10^{-7}, eight warmup steps, weight decay 0.01, KL coefficient 0.001, and clipping bounds 0.20/0.28. Checkpoint 419 is the zero-indexed final iteration of this schedule, corresponding to approximately one pass through the training corpus. We run matched seeds 42, 43, and 44. The optimization difference is $A_{i,t} = A_i$ for B4 and Equation 8 for SLR.

4.4 Rewards and metrics

Training combines the sentence-level xCOMET-XL score (Rei et al., 2020; Guerreiro et al., 2024), sentence chrF++ with word order 2, and sentence BLEU, combined as Equation 1. BLEU and chrF++ use SacreBLEU 2.6.0 (Papineni et al., 2002; Popović, 2017; Post, 2018).

GlotLID-v3 predicts `ind_Latn` for 126/1,012 FLORES+ Malay references (12.5%) and 432/1,997 NTREX Malay references (21.6%), or 558/3,009 (18.54%) pooled. This disagreement on known-Malay references motivates a graded language penalty instead of hard rejection (Kargaran et al., 2023; Nomoto, 2023). For Malay targets, $P_{\text{lid}} = 0$ for `zsm_Latn`/`zsm_Arab` and $P_{\text{lid}} = 0.50 + 0.30c$ otherwise, including `ind_Latn`, where c is GlotLID confidence. If ρ_{ws} is the proportion of nonpunctuation characters in disallowed script ranges, then $P_{\text{script}} = \min(1, 3\rho_{\text{ws}})$. For Chinese targets, $P_{\text{lid}} = 0$ and P_{script} is 0, 0.3, or 0.8 for Han ratios $h \geq 0.45$, $0.20 \leq h < 0.45$, or $h < 0.20$, respectively.

We report every reward component, the SoftLID-adjusted score, and a separate hard-evaluation validity flag: GlotLID must return an accepted Malay or Sinitic target label. We also score all checkpoint-419 outputs with the reference-based MetricX-24 Hybrid XXL v2p6 model; lower scores are better on its 0–25 scale (Juraska et al., 2024). MetricX is not used during training.

4.5 Human evaluation

The primary evaluation contains 300 unique Chinese-to-Malay sources at checkpoint 419, balanced over the two datasets, three seeds, and two prompt languages. Each item presents the Chinese source and two system outputs. System identity, the Malay reference, metric and reward values, LID decisions, and automatic error spans are hidden. Output order is randomized and balanced 150/150.

The author, who is bilingual in Chinese and Malay, recorded pairwise adequacy and fluency preference (A/B/tie), a separate variety label for each output (Malaysian Standard Malay, Indonesian, mixed, other, or uncertain), and whether each output contains an omission or major error. The protocol separates error category from severity and treats omission as an accuracy error, following MQM-style analysis without implementing full MQM (Freitag et al., 2021).

The author separately audited span localization on 150 paired Chinese-to-Malay outputs: 100 randomly sampled candidates form the primary set and 50 disagreement-enriched candidates are diagnostic. The same author-evaluator marked target error spans and source omissions in two separate passes from blank templates. Each pass displayed the Chinese source, Malay reference, and one candidate output, while system identity, metric and reward values, LID decisions, and xCOMET spans remained hidden. The second pass did not expose the first-pass labels. The repeated passes quantify intra-annotator repeatability, and the primary span reference is the intersection of target words marked in both passes.

4.6 Statistical analysis

We pair B4 and SLR outputs for every seed, dataset, direction, source, and prompt language. The primary cluster is a unique source–direction pair; seeds and prompt variants remain repeated observations within a cluster. System means first average within source. Confidence intervals use 20,000 paired bootstrap resamples of the 6,018 unique sources (36,108 output pairs). Common-valid results condition on both systems passing LID and are diagnostic because validity is affected by the policy.

Human preference effects are the proportion preferring SLR minus the proportion preferring B4, with ties retained. Error effects are paired differences in event rates. We use 50,000 paired item-bootstrap resamples and two-sided exact sign tests. Span-localization intervals use 10,000 source-clustered bootstrap resamples over the 100 unbiased candidates.

5 Automatic Results

5.1 The optimized neural signal improves

SLR increases the raw MT reward by 0.0021 (95% source-clustered CI $[0.0013, 0.0030]$; Table 1). COMET accounts for nearly all of the gain: $+0.0046$ overall and $+0.0080$ for Chinese-to-Malay. Overall chrF++ and BLEU intervals cross zero, while Malay-to-Chinese chrF++ decreases by 0.0015. Because COMET supplies half of the reward and xCOMET supplies the local spans, its increase measures alignment with the optimized signal.

5.2 MetricX favors SLR, primarily on NTREX

MetricX-24 favors SLR by -0.082 points overall on its 0–25 scale, where lower is better (Table 2). The per-seed differences are -0.089, -0.096, and -0.061 for seeds 42–44. The gain is larger for Chinese-to-Malay (-0.118) than Malay-to-Chinese (-0.046).

The effect is concentrated on NTREX. Its largest slice is NTREX Chinese-to-Malay (-0.138; 95% CI $[-0.201, -0.078]$). The corresponding FLORES+ direction-specific estimates, -0.029 for Malay-to-Chinese and -0.078 for Chinese-to-Malay, both have intervals crossing zero. MetricX thus favors SLR on this aggregate comparison, while its direction-level FLORES+ estimates remain unresolved.

5.3 Language control offsets the semantic gain

The SoftLID-adjusted score decreases by 0.0038 overall. Target-language validity falls from 93.16% to 92.32%, driven by Chinese-to-Malay (86.78% to 84.95%); Malay-to-Chinese validity rises by 0.16 points. The human variety labels in Section 6 do not show a corresponding increase in Indonesian/mixed output, so detector validity and human variety are reported as distinct outcomes.

5.4 Common-valid outputs retain the metric pattern

Among rows where both systems pass the language detector, MetricX still favors SLR (-0.079; 95% CI $[-0.110, -0.048]$) and COMET increases by 0.0044 (Table 3); chrF++ and BLEU remain unresolved. The learned-metric gains therefore persist beyond the additional LID failures. This subset is diagnostic because validity is affected by the policy.

5.5 Local credit is active but rarely changes sign

Across 120,960 training responses, 77.54% receive a nonzero local residual. The broad-span gate removes 12.96%, no detected span removes 6.82%, and invalid UTF-8 removes 0.10%. The cap is reached on 1.76% of all samples; $|\lambda g| \leq 0.5$ and the maximum absolute zero-sum error is 4.3×10^{-14}. A word-level sign-change event occurs in 11.90% of samples. Most active residuals therefore adjust magnitude without reversing a token's sequence-level credit.

6 Human Evaluation and Span Actionability

6.1 Translation quality

Adequacy yields 121 SLR preferences, 134 B4 preferences, and 45 ties; the -4.33 point interval crosses zero. Fluency is likewise unresolved (128/147/25; -6.33 points). The author's bilingual judgments therefore do not confirm the learned-metric gain as an adequacy or fluency improvement.

Omission differs clearly. The author marked omissions in 9.33% of B4 outputs and 16.67% of SLR outputs, a $+7.33$ point increase with 95% CI $[+2.33, +12.33]$. Major-error and Indonesian/mixed-variety differences remain unresolved. In particular, human variety labels do not support treating every GlotLID rejection as Indonesian drift.

6.2 Learned metrics and human preferences

On the 300-item sample, the mean MetricX difference is $+0.125$ after orienting positive values toward SLR, with a confidence interval crossing zero. MetricX and COMET agree with the non-tied human adequacy winner on 67.7% and 70.1% of eligible items (Table 6). The aggregate disagreement comes partly from effect magnitude: on human-SLR wins, mean MetricX and COMET improvements are $+1.330$ and $+0.0781$; on human-B4 wins, they are -1.048 and -0.0266. Continuous metric averages can therefore favor SLR even when B4 receives more pairwise votes.

Omission reveals a construct mismatch beyond this arithmetic. Among 42 items where only SLR is marked omissive, adequacy favors B4/SLR/tie on 27/11/4, yet mean MetricX and COMET still

Scope	Metric (↑)	B4	SLR	SLR−B4 [95% CI]
Overall	SoftLID score	0.4668	0.4631	-0.0038 [-0.0063, -0.0012]
Overall	R_{MT}	0.5186	0.5208	+0.0021 [+0.0013, +0.0030]
Overall	COMET	0.7618	0.7663	+0.0046 [+0.0033, +0.0058]
Overall	chrF++	0.3277	0.3272	-0.0004 [-0.0013, +0.0004]
Overall	BLEU	0.1972	0.1972	+0.0000 [-0.0009, +0.0009]
Overall	Target-language valid (%)	93.16	92.32	-0.84 [-1.17, -0.51]
ms→zh	SoftLID score	0.5038	0.5051	+0.0014 [+0.0002, +0.0026]
ms→zh	R_{MT}	0.5086	0.5086	-0.0000 [-0.0010, +0.0010]
ms→zh	COMET	0.7681	0.7692	+0.0011 [-0.0002, +0.0024]
ms→zh	chrF++	0.2311	0.2296	-0.0015 [-0.0027, -0.0004]
ms→zh	BLEU	0.2759	0.2755	-0.0004 [-0.0019, +0.0010]
ms→zh	Target-language valid (%)	99.55	99.70	+0.16 [+0.08, +0.24]
zh→ms	SoftLID score	0.4299	0.4210	-0.0090 [-0.0140, -0.0041]
zh→ms	R_{MT}	0.5287	0.5330	+0.0043 [+0.0030, +0.0056]
zh→ms	COMET	0.7554	0.7634	+0.0080 [+0.0058, +0.0102]
zh→ms	chrF++	0.4243	0.4249	+0.0006 [-0.0007, +0.0018]
zh→ms	BLEU	0.1185	0.1189	+0.0005 [-0.0007, +0.0017]
zh→ms	Target-language valid (%)	86.78	84.95	-1.84 [-2.49, -1.19]

Table 1: Checkpoint-419 automatic evaluation over three matched seeds. Means are source-weighted; 95% intervals resample unique sources while keeping seeds and prompt-language variants within cluster. All metrics are higher-better. The SoftLID score includes the language and script penalties; R_{MT} does not. Validity differences are percentage points.

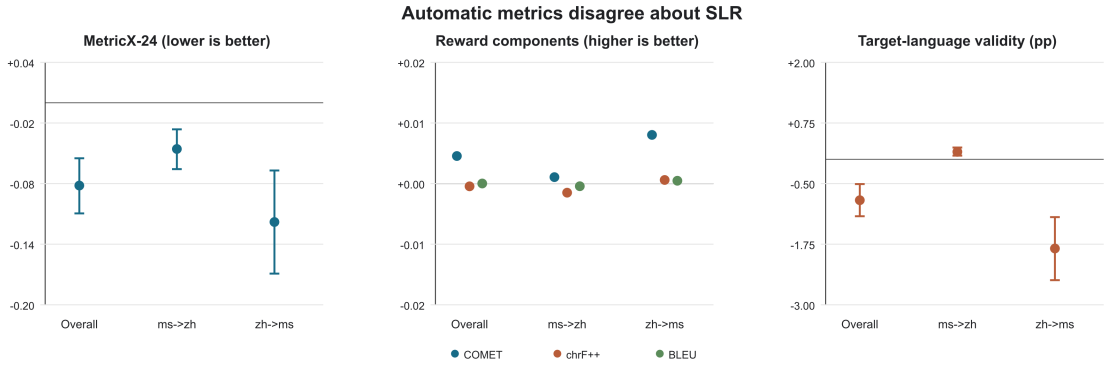

Figure 2: Checkpoint-419 effects. MetricX-24 and COMET favor SLR, lexical metrics remain near zero, and Chinese-to-Malay target-language validity decreases. Error bars show 95% unique-source bootstrap intervals where available.

Scope	B4	SLR	SLR−B4 [95% CI]
Overall	3.992	3.910	-0.082 [-0.110, -0.055]
ms→zh	3.508	3.462	-0.046 [-0.066, -0.026]
zh→ms	4.476	4.358	-0.118 [-0.169, -0.067]
FLORES+	3.362	3.309	-0.053 [-0.102, -0.006]
NTREX-128	4.311	4.215	-0.096 [-0.130, -0.063]

Table 2: Reference-based MetricX-24 scores at checkpoint 419 (↓). Negative differences favor SLR. Intervals use unique-source clusters spanning seeds and prompt variants.

favor SLR ($+0.268$ and $+0.0046$). chrF++ and BLEU favor B4 (-0.0245 and -0.0158). MetricX and COMET differences have near-zero rank correlation with the paired omission label ($+0.053$ and -0.061). Within these cases SLR is 1.62 words shorter on average, although its overall length differs from B4 by only $+0.12$ words. On this subset, favorable learned-metric changes coexist with worse human coverage and do not encode the omission penalty expressed by the author-evaluator.

COMET supplies half of the training reward and xCOMET supplies the spans. MetricX is not used in training, but its item-level difference correlates with COMET at $r = 0.750$, and their winner signs agree on 71.8% of common-valid nonties. Their agreement provides convergent learned-semantic evidence.

Scope	Metric	Sources	B4	SLR	SLR−B4 [95% CI]
Overall	MetricX-24 (↓)	5776	3.997	3.918	-0.079 [-0.110, -0.048]
Overall	R_{MT} (↑)	5776	0.5170	0.5190	+0.0020 [+0.0011, +0.0029]
Overall	COMET (↑)	5776	0.7592	0.7635	+0.0044 [+0.0029, +0.0058]
Overall	chrF++ (↑)	5776	0.3248	0.3244	-0.0005 [-0.0014, +0.0005]
Overall	BLEU (↑)	5776	0.1997	0.1995	-0.0001 [-0.0012, +0.0009]
ms→zh	MetricX-24 (↓)	3005	3.506	3.461	-0.045 [-0.066, -0.025]
ms→zh	R_{MT} (↑)	3005	0.5087	0.5087	+0.0000 [-0.0010, +0.0010]
ms→zh	COMET (↑)	3005	0.7684	0.7694	+0.0011 [-0.0002, +0.0023]
ms→zh	chrF++ (↑)	3005	0.2309	0.2295	-0.0014 [-0.0026, -0.0003]
ms→zh	BLEU (↑)	3005	0.2760	0.2757	-0.0004 [-0.0019, +0.0010]
zh→ms	MetricX-24 (↓)	2771	4.530	4.415	-0.115 [-0.178, -0.054]
zh→ms	R_{MT} (↑)	2771	0.5259	0.5301	+0.0042 [+0.0026, +0.0058]
zh→ms	COMET (↑)	2771	0.7492	0.7571	+0.0080 [+0.0054, +0.0106]
zh→ms	chrF++ (↑)	2771	0.4267	0.4273	+0.0006 [-0.0009, +0.0020]
zh→ms	BLEU (↑)	2771	0.1168	0.1170	+0.0001 [-0.0013, +0.0016]

Table 3: Checkpoint-419 common-valid-language diagnostic over three matched seeds. Rows are retained only when both systems pass the target-language detector. Means are source-weighted; intervals resample unique sources while keeping seed and prompt variants within cluster. Because validity is a post-treatment outcome, these estimates are diagnostic rather than primary.

SLR outcome	Samples	Rate
Active local residual	93,794	77.54%
No spans	8,244	6.82%
No mapped span	0	0.00%
Uniform lexical quality	1,383	1.14%
Broad-span gate	15,674	12.96%
Wrong-script gate	1,748	1.45%
Invalid-UTF8 gate	117	0.10%

Table 4: Training-time SLR activation and reliability gates, pooled over all three seeds. Gated samples use the unmodified sequence-level advantage.

Measure	B4	SLR	Δ [95% CI]
Adequacy preference	44.67	40.33	-4.33 [-14.67, +6.00]
Fluency preference	49.00	42.67	-6.33 [-17.00, +4.67]
Omission rate	9.33	16.67	+7.33 [+2.33, +12.33]
Major-error rate	53.67	56.00	+2.33 [-6.67, +11.33]
Indonesian/mixed rate	11.67	9.67	-2.00 [-5.67, +1.67]

Table 5: Blinded evaluation by the bilingual author on 300 Chinese-to-Malay sources. All entries are percentages; Δ is SLR minus B4 in percentage points. Preference rates retain 45 adequacy and 25 fluency ties. Intervals use 50,000 paired item-bootstrap resamples.

6.3 Error detection, localization, and residual orientation

Across the author's two blinded passes, intra-annotator repeatability for sentence-error presence is 91% for the unbiased candidates (Cohen's $\kappa = 0.591$), omission repeatability is 93% ($\kappa = 0.760$), and target-word macro F1 is 0.700. This repeatability provides a stability context for the detector comparison.

xCOMET identifies most sentences marked erroneous in both passes: any-span precision is 0.824 and recall is 0.904. Localization is much weaker. Macro target-word F1 is 0.259 (95% source-clustered CI $[0.200, 0.322]$), and strict character IoU is 0.177. After applying the same lexical gates used during training, the mean fraction of target words marked erroneous in both passes that receive a negative residual is 0.450 (95% CI $[0.335, 0.572]$; micro count 100/228). The residual direction agrees with the repeated-pass reference on fewer than half of the marked error words.

Localization quality does not differ by system: macro F1 is 0.270 for B4 and 0.249 for SLR, a paired difference of -0.020 with 95% CI $[-0.107, 0.071]$. Omissions expose a separate structural gap. An omitted source phrase has no generated target token, so the current target-span mapper cannot assign direct insertion credit to it.

7 Discussion

SLR optimizes a learned semantic signal. The residual is numerically stable and moves both reward-aligned COMET and MetricX, a learned metric not used in training. The blinded evaluation by the bilingual author does not resolve an adequacy or fluency preference and identifies a higher omission rate; chrF++ and BLEU add no resolved lexical evidence. The MetricX improvement is strongest on reciprocal NTREX and is unresolved within each FLORES+ direction. Together, these results characterize SLR as metric-specific policy change rather than an overall improvement in Chinese–Malay translation.

Metric	n	Mean SLR improvement [95% CI]	Adequacy winner agreement (%)	ρ with omission difference
MetricX-24	300	+0.1251 [-0.2278, +0.4720]	67.7	+0.053
COMET	244	+0.0195 [+0.0040, +0.0349]	70.1	-0.061
chrF++	300	+0.0059 [-0.0122, +0.0239]	55.6	-0.230
BLEU	300	+0.0064 [-0.0016, +0.0148]	53.6	-0.139

Table 6: Automatic metrics on the 300-item human-evaluation sample. Positive values favor SLR; MetricX is reoriented as B4 minus SLR because lower is better. COMET includes only the 244 pairs for which both systems pass the language gate. Winner agreement excludes human and metric ties. Correlations are descriptive Spearman coefficients.

Measure	Estimate	95% CI
Any-span precision / recall	0.824 / 0.904	–
Macro word precision / recall	0.208 / 0.509	–
Macro word F1	0.259	[0.200, 0.322]
Error words with negative residual	0.450	[0.335, 0.572]

Table 7: xCOMET localization and gated word-residual orientation on the 100 unbiased Chinese-to-Malay candidates. The reference is the target-word intersection of two blinded passes by the same bilingual author. Intervals use 10,000 source-clustered bootstrap resamples. The residual row is a word-level orientation diagnostic, not an observed final-advantage sign flip.

Language control is a second measurement axis. GlotLID validity decreases for Chinese-to-Malay, while the human labels do not show more Indonesian/mixed output. GlotLID's 18.54% Indonesian-prediction rate on known-Malay references shows why classifier validity and human variety must remain separate outcomes.

Detection is not actionable credit. The span audit identifies an actionability gap in the training signal. xCOMET usually detects that an erroneous sentence contains an error, but its word boundaries align poorly with the two-pass annotation. Reliability gates remove the broadest and least informative cases, yet the remaining residual places negative credit on only 45% of error words marked consistently across the two passes. A computable local residual is therefore not necessarily a correctly directed policy update. The gap between activation and credit reversal is also visible in the training diagnostics: 77.5% residual activation produces a word-level sign change in only 11.9% of responses.

Omission is harder still. Target-side spans can redistribute credit only over generated content. They cannot identify a missing phrase or specify where the policy should insert it. Source-coverage states, EOS credit, and explicit insertion or edit actions are candidate representations for this error class.

Relation to successful fine-grained MT optimization. Ramos et al. (2026) optimize local severity values directly with token-level REINFORCE/PPO on English–German and English–French. Finkelstein et al. (2025) combine MetricX with a Gemma-based AutoMQM evaluator at multilingual scale. These methods differ in evaluator, language pair, and optimization geometry. Their gains establish that local MT feedback can work; our results show that the evaluator-to-policy mapping is a separate empirical requirement. Fine-grained MT-RL should validate span localization, effective credit activation, and whether the representation covers the required correction, while reporting reward components, language control, and human error outcomes in addition to aggregate reward.

8 Conclusion

We tested whether xCOMET error spans provide useful signed policy credit for Chinese–Malay GRPO. SLR improves COMET and MetricX-24 across three matched seeds, while chrF++ and BLEU differences are unresolved; it lowers the SoftLID-adjusted score and increases human-marked omission. The author's 300-item blinded bilingual evaluation does not resolve an adequacy or fluency preference.

The repeated span audit exposes a credit-alignment gap. xCOMET reaches 0.904 sentence-level error recall but only 0.259 macro target-word F1, and the gated residual assigns negative credit to 45.0% of target words marked erroneous in both passes. Error detection, localization, and policy-credit actionability are therefore distinct properties. Before their spans are used as token-level rewards, explainable metrics should be evaluated for detection, localization, and policy actionability, including whether the credit representation can ex-

press the required correction.

Limitations

The training comparison covers one 9B model, one language pair, one residual strength, and xCOMET-XL as the local evaluator. NTREX is English-sourced; reciprocal Chinese–Malay evaluation therefore contains translationese and occasional reference errors. COMET is part of the reward, MetricX is another learned metric, and GlotLID is imperfect for Malay–Indonesian separation.

The single author, who is bilingual in Chinese and Malay, completed the quality evaluation and both span-audit passes. System identity and experimental signals were withheld, and A/B order in the quality evaluation was randomized and balanced. The results may nevertheless reflect evaluator-specific judgments, and no independent second evaluator was available; the span-pass statistics measure intra-annotator repeatability rather than inter-annotator validation. The span audit measures word-level residual orientation rather than replaying final policy-token advantages. SLR also cannot assign direct insertion credit to content omitted from the generated response.

Ethics Statement

The target construct is Malaysian Standard Malay, not a claim that Indonesian is linguistically inferior. Malay and Indonesian are closely related standard varieties associated with different speech communities and national contexts (Nomoto, 2023). Variety drift is a task mismatch only because the requested and referenced target is Malaysian Standard Malay.

The author, who is bilingual in Chinese and Malay, voluntarily completed the 300-item evaluation and both span-audit passes under the blinding protocols described in the experimental setup. No external participant was recruited. The internal training corpus, including its raw web text and cleaned parallel pairs, is not redistributed.

Acknowledgments

The author thanks Bojie Wang for guidance in setting up the experimental environment, advice on tokenization, and for proposing the initial token-level GRPO direction explored in this project. The author also thanks Menyang Sun for identifying and assembling the evaluation datasets and benchmarks, and for collecting the Malay-language web data used to construct the training corpus. The author is grateful to Tsinghua University's Student Research Training (SRT) program for the opportunity and support that enabled this project, and thanks the program's mentors for helpful discussions and guidance throughout.

Generative-AI assistance. The author used OpenAI's ChatGPT and Codex, as well as Anthropic's Claude Code, for code development and debugging, literature search, analysis support, figure preparation, and manuscript drafting and revision. All scientific claims, interpretations, and reported results were verified by the author.

References

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24. Association for Computational Linguistics.

Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. OpenLID-v3: Improving the precision of closely related language identification – an experience report. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 275–292, Rabat, Morocco. Association for Computational Linguistics.

Zhaopeng Feng, Jiahan Ren, Jiayuan Su, Jiamei Zheng, Hongwei Wang, and Zuozhu Liu. 2025. MT-RewardTree: A comprehensive framework for advancing LLM-based machine translation via reward modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18556–18567, Suzhou, China. Association for Computational Linguistics.

Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Markus Freitag, and David Vilar. 2025. Google Translate's research submission to WMT2025. In *Proceedings of the Tenth Conference on Machine Translation*, pages 723–731. Association for Computational Linguistics.

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.

Javier Garcia Gilabert, Carlos Escolano, Xixian Liao, and Maite Melero. 2025. Terminology-constrained

translation from monolingual data using GRPO. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1335–1343, Suzhou, China. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218. Association for Computational Linguistics.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. 2026. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24681–24693, San Diego, California, United States. Association for Computational Linguistics.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Hiroki Nomoto. 2023. Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 76–82. 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. 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. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André F. T. Martins. 2026. Fine-grained reward optimization for machine translation using error severity mappings. *Transactions of the Association for Computational Linguistics*, 14:733–754.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Yunzhi Shen, Hao Zhou, Xin Huang, Xue Han, Junlan Feng, and Shujian Huang. 2026. PEGRL: Improving machine translation by post-editing guided reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17225–17242. Association for Computational Linguistics.

Zeli Su, Ziyin Zhang, Zhou Liu, Xuexian Song, Zhankai Xu, Longfei Zheng, Xiaolu Zhang, Rong Fu, Guixian Xu, and Wentao Zhang. 2026. Reinforcement learning with semantic rewards enables low-resource language expansion without alignment tax. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17772–17786, San Diego, California, United States. Association for Computational Linguistics.

Team GLM. 2024. ChatGLM: A family of large language models from GLM-130b to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*.

Klaudia Thellmann, Bernhard Stadler, Michael Färber, and Jens Lehmann. 2026. Quantifying the impact of translation errors on multilingual LLM evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 41311–41334. Association for Computational Linguistics.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. 2026. SSR-zero: Simple self-rewarding reinforcement learning for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 6039–6052, San Diego, California, United States. Association for Computational Linguistics.

A Implementation Checks

For every active sample, the implementation enforces

$$\left|\sum_{t\in\mathcal{T}_i} g_{i,t}\right| < 10^{-8}, \tag{10}$$

$$\max_t |g_{i,t}| \le 2, \tag{11}$$

$$g_{i,t} = 0 \text{ for masked tokens}, \tag{12}$$

$$A_{i,t}^{SLR}\big|_{\lambda=0} = A_i. \tag{13}$$

Across the three runs, the largest zero-sum residual is 4.3×10^{-14}. Every generated token ID has one byte-level map entry. Invalid UTF-8 is the only mapping exception assigned $g = 0$; all other response, backend, and token-count mismatches raise an error.

B Additional MetricX Analysis

The three seed-level MetricX differences are -0.0894, -0.0956, and -0.0608. A Student-t interval over the three seed means is $[-0.1281, -0.0358]$; a hierarchical sensitivity bootstrap that resamples both seeds and sources gives $[-0.1209, -0.0414]$. The primary estimate remains the unique-source clustered interval in Table 2.

COMET and MetricX contain 2,469 and 2,436 exact paired ties. Excluding pairs tied under either metric, their preference direction agrees on 20,105/29,542 pairs (68.1%).

C Data and Annotation Checks

The final 4,691-row training corpus contains no exact normalized sentence match with either evaluation suite. Four FLORES+ and six NTREX unique pairs share at least one 8-gram with training data; all matches occur on the Chinese side.

For the author's two blinded span-audit passes, intra-annotator sentence-error agreement is 0.91 ($\kappa = 0.591$), omission agreement is 0.93 ($\kappa = 0.760$), target-word macro F1 is 0.700, and target-character IoU is 0.623. The primary reference uses the intersection of target words marked in both passes. The diagnostic 50-candidate stratum is excluded from all primary span estimates.