QEbreak at WMT26: A Pre-Registered Audit of Omission and Addition Asymmetry in Machine Translation Quality Metrics

Maksim Silchenko
National University of Singapore
maks.silchenko.04@gmail.com

Abstract

Reference-free quality estimation (QE) metrics see only the source and the hypothesis, so a translation that silently drops source content leaves nothing for them to flag. QEbreak, a pre-registered contrastive challenge set submitted to the WMT26 Automated Translation Quality Evaluation task (Subtask 4), measures this blind spot: 2,886 segments over 10 language directions, built as matched {base, omission, addition} triples, a fact-free verbosity ladder, and numeric-contradiction pairs as a positive control. Every hypothesis, endpoint, and equivalence margin was frozen in git-timestamped commits before any score existed. Scores from 31 systems confirm the deficit and refute our mechanism. Reference-free QE metrics under-penalize omission (family asymmetry $A = -0.51$; a current QE baseline scores the omission at or above its own base in 19.5 percent of matched contests, strictly above in 14.3), and the edit-direction by reference-availability interaction is significant ($\beta = -0.069$, 95% CI $[-0.081, -0.057]$). But reference-based neural metrics are asymmetric in the same direction ($A = -0.42$), outside the pre-registered ± 0.30 equivalence margin, so omission under-penalization belongs to learned metrics as a class; reference access softens it but does not remove it. Every returned metric penalizes pure verbosity, and two systems score corrupted numerals above correct ones.

1 Introduction

A quality estimation metric that scores translations without a reference is the natural reward signal for machine translation systems: it is cheap, it needs no human reference, and it is already used to rerank candidate outputs. That role raises a structural question. A reference-based metric can notice that translated content is missing because the reference still contains it. A reference-free metric has only the source and the hypothesis, and a hypothesis that omits content is shorter, more fluent, and internally consistent. If reference-free metrics penalize omission less than they penalize a comparable amount of added content, then a best-of-n reranker built on such a metric will sometimes prefer the incomplete translation. QEbreak was built to measure that preference directly.

QEbreak tests three pre-registered hypotheses. The headline hypothesis $\mathrm{H_B}$: reference-free QE metrics under-penalize meaning omission relative to fluent addition, while reference-based metrics penalize the two edit directions roughly symmetrically. $\mathrm{H_A}$: reference-free QE scores do not decrease, or even increase, as fluent, fact-free, source-unsupported material is appended to a realistic base translation. $\mathrm{H_C}$, a positive control: the reference-based neural frontier catches a single load-bearing numeric contradiction while surface and reference-free QE metrics diverge, which proves the harness can detect breaks at all.

The design choice that separates this audit from prior challenge sets is that every claim was frozen before any score existed. The hypotheses, endpoints, equivalence margin, multiplicity correction, provenance strata, and orientation manifest were committed to a timestamped git history before the challenge set was uploaded, and the three pre-score amendments are themselves logged commits (Appendix A). Whatever the organizer scores say, the reported outcome is the pre-registered one: a confirmation, or an honest negative.

We contribute three things. First, a 2,886-segment contrastive challenge set over 10 language directions, built as matched same-source triples with controlled information-unit edits and submitted blind with opaque identifiers and a private keymap. Second, a pre-registered analysis with a signed interaction primary endpoint, a reward preference-flip rate that directly measures the reranking failure, and a numeric equivalence

anchor for the reference-based family. Third, the findings themselves: a significant reference-availability interaction on top of an asymmetry that turns out to be shared by every learned family, a preference-flip rate that reaches one matched contest in five for a current QE baseline, a verbosity hypothesis refuted outright, and two returned systems that score corrupted numerals above correct ones.

2 Related Work

ACES (Amrhein et al., 2022) and its span-annotated extension (Moghe et al., 2025) established the contrastive challenge-set paradigm, including single-edit omission and addition items. `QEbreak` differs on three axes: the omission and the addition are matched on the same source at a controlled information-unit dose, the contrast is carried through the reference-availability axis, and the analysis reaches the reward-model preference flip that measures the deployment-relevant event directly. The MT-breaker line of adversarially difficult inputs (Zouhar et al., 2026) sets the rigor bar we adopt. On the length axis, prior work diagnosed QE length bias (Zhang et al., 2025) and a WMT25 challenge set measured continuation drift (CoDrift, in Lavie et al., 2025); Arm A instead isolates pure, fact-free, unsupported length as a dose curve. The pre-registered family taxonomy covers surface metrics (Papineni et al., 2002; Popović, 2015; Snover et al., 2006), reference-based neural metrics (Rei et al., 2022a; Sellam et al., 2020; Guerreiro et al., 2024; Juraska et al., 2024), reference-free QE (Rei et al., 2022b), and LLM judges in the GEMBA family (Kocmi and Federmann, 2023).

3 The QEbreak Challenge Set

`QEbreak` was submitted to WMT26 Subtask 4: 2,886 segments across 10 language directions, 231,582 tokens under the organizer counting rule, well under the 1M-token cap; the organizers confirmed receipt on 22 July 2026. Table 1 gives the composition.

Direction	A	B	C	Total
en-de	330	300	264	894
en-cs	0	219	210	429
cs-uk	0	261	156	417
en-ru	126	45	192	363
en-zh	0	66	188	254
cs-de	0	0	216	216
en-ar	0	0	176	176
en-ko	0	0	66	66
en-ja	0	33	18	51
zh-ja	0	0	20	20
Total	456	924	1,506	2,886

Table 1: Shipped segments per language direction and arm. Arm A: 76 six-step verbosity dose series (456 segments). Arm B: 308 matched {base, omission, addition} triples (924 segments). Arm C: 753 numeric minimal pairs (1,506 segments).

3.1 Three arms

Arm B (headline): matched omission and addition triples. Each of the 308 triples holds the source fixed and pairs a base translation with an omission that removes one information unit and an addition that appends one fact-free, source-unsupported unit. An information unit is one span removed or added whole and verified by the gates of Section 3.2. In the deterministic strata (S2 and S3 of Section 3.3) it is a trailing sentence or clause for an omission, and for an addition one reference segment of another source in the same language direction, appended whole and gate-verified fact-free. In the LLM-generated stratum (S1) it is one small factual detail for an omission (a modifier, name, number, date, or short phrase) and one short generic clause or sentence for an addition. Omission and addition are the two edit directions of the design. Members travel as separate rows paired by opaque group identifiers, matching the organizer-endorsed separate-segments design. The triple supports both primary endpoints: the signed edit-direction by reference-availability interaction (N1) and the reward preference-flip rate P_{flip}, the fraction of matched {base, omission} pairs (the contests a reranker adjudicates) in which a reference-free metric scores the omission at or above the base (N2). The two edit directions are matched on information units, not edit size: additions change a median of 2.0 times as many characters as omissions remove (whitespace-stripped characters, the unit used for the edit-size ratios and the magnitude covariate), so a length-controlled sensitivity refit accompanies the primary interaction (Section 4).

Arm A (secondary): a verbosity dose ladder. Each of the 76 series appends 1 to 5 fluent, fact-free filler units to a base translation on en-de and en-ru, giving six segments per series including the dose-0 base. Fillers are screened to contain no digits, dates, or named entities, and to be entailment-

neutral against both the source and the base (Section 3.2), so the ladder isolates pure unsupported length. At the maximum dose the hypothesis grows to a median of 6.4 times the base length.

Arm C (positive control): numeric minimal pairs. Each of the 753 pairs contrasts a correct hypothesis with a copy whose single load-bearing numeral is corrupted: digit substitution (440 pairs), transposition (213), or a magnitude shift (100). Numeral selection follows a scoring rule frozen in the pre-registration (units and separators weighted up, years penalized and never corrupted as a fallback), so the corrupted numeral is never a researcher degree of freedom. Arm C spans all 10 directions and proves the harness has teeth: if the reference-based frontier misses a flat numeric contradiction, the harness, not the metrics, is under suspicion.

3.2 Validity gates

Every generated item passes deterministic validity gates before it can ship; items that fail are dropped, not repaired. The omission gate requires the base hypothesis, the stem the omission was cut from (the reference itself in stratum S3 below), to entail the omitted-member hypothesis while the dropped unit stays un-entailed by the remaining text, using a natural language inference (NLI) model (mDeBERTa-v3-base-xnli, Laurer et al., 2024), which blocks omissions that merely rephrase. The addition gate requires the appended unit to carry no digit, date, or named entity (XLM-R named-entity recognition, NER) and to be entailment-neutral against both source and base (neutral ≥ 0.70, entailment and contradiction ≤ 0.15 each). It validates the appended delta itself whenever the addition is a clean append of the base, the base ending at a word boundary (every deterministic addition and 70 of the 89 LLM additions); for the other 19 (18 where the addition is not a verbatim extension of the base, and one en-zh append whose base ends without a boundary the rule recognizes) it checks the unit the generator itself reported as appended and requires the enlarged hypothesis to entail the base, since in those cases the base may carry small edits. Subtle LLM-generated triples obey symmetric length caps (at most 25 percent of the base's characters dropped or grown, enforced on raw character counts; realized worst cases 22.9 and 24.4 percent), and all perturbed members must be distinct from their base and from each other.

Gate survival is low by design. In the LLM-generation runs, 1,042 raw candidates were drawn, 861 parsed into complete triples, and the gates admitted 89 of those: 10.3 percent overall, between 3.2 percent (en-ja, 4 of 125) and 20.0 percent (en-ru, 6 of 30) per direction, with the full counts in Section 4. No gate reads any metric score.

3.3 Provenance strata

Arm B ships in three pre-declared strata, defined purely from metadata written at generation time. S1 (89 triples; the registration calls it the powered stratum): subtle perturbations generated by Llama-3.1-8B on real WMT24 system outputs under the length caps; the omission typically drops a single detail such as a street number. S2 (100 triples): deterministic whole-unit edits on real WMT24 system-output bases. S3 (119 triples): deterministic edits on reference-verbatim bases. Real-MT bases come from the median-ranked WMT24 system per direction (Kocmi et al., 2024), watermark canary lines filtered out. The pre-registered secondary prediction is a gradient, for the reference-free QE family, in the relative asymmetry A of Section 4 (more negative means omission is penalized less, relative to addition) and in P_{flip}: $A(\text{S1}) < A(\text{S2} + \text{S3})$ and $P_{\text{flip}}(\text{S1}) > P_{\text{flip}}(\text{S2} + \text{S3})$, no prediction between S2 and S3, and no stratified test in the confirmatory family. The rationale was recorded before any score existed: a local CometKiwi pre-filter, a stand-in for the audit-target family, labeled every item with a break prediction (a prediction that a metric would score the perturbed member at or above its base; Section 4), and it already penalizes the crude deterministic omissions, so pooled dilution is a described property of the design, not a post-hoc rescue.

3.4 Blind submission protocol

The uploaded rows carry exactly the seven published fields; all internal metadata was stripped, identifiers are opaque hashes with no label, phenomenon, or dose token, and row order was decorrelated so the base member sits at no fixed position (base-first fell from 100 percent to 44.0 percent, against a 43.3 percent chance level). A private keymap, never uploaded, restores the pairing when scores return. Residual structure inherent to contrastive sets, such as pair members sharing a source, is disclosed in the Limitations.

4 Pre-Registered Analysis Plan

The full pre-registration lives in a timestamped git history frozen before the challenge-set upload and before any organizer score existed. Appendix A records the commits, the three pre-score amendments, and the post-return changes to the analysis code, which were made in one pre-analysis commit, include changes the registration did not anticipate, and are reported there as deviations. Table 2 states every registered quantity, its decision rule, its predicted sign, and where Section 5 reports it; the paragraphs below define the terms the table uses, give the anti-circularity counts, and name the two sensitivity refits, one registered and one added after the return.

The N1 model, as fitted. For metric m and each perturbed Arm B member, the signed drop is the oriented base score minus the oriented member score, so a positive value means the metric penalized the edit; oriented means sign-adjusted so that higher is better, and since every system on this return already reported higher-is-better scores, orientation changed no sign here and the manifest's role is to fix each metric's family and reference indicator. Within each metric the drops are replaced by their average rank scaled to $(0, 1)$, so the response y is a within-metric percentile and is invariant to any monotone rescaling of a metric (rank-normalization is the registered primary transform; a z-score refit is the registered sensitivity analysis). The registered three-way design (edit direction, reference availability, dose) reduces, because every shipped Arm B edit carries one information unit, to the identifiable model

$$y = \beta_0 + \beta_1 D_{\text{add}} + \beta_2 R + \beta\,(D_{\text{add}} \cdot R) + \varepsilon, \quad (1)$$

where $D_{\text{add}} = 1$ for an addition member (omission is the baseline) and $R = 1$ when the metric reads the reference. R is set by the manifest, not by family: $R = 0$ pools the four reference-free QE metrics and the two reference-free LLM judges (6 metrics, 3,444 observations); $R = 1$ pools the five reference-based neural metrics, the three surface metrics, and the three reference-reading LLM judges (11 metrics, 6,776 observations). The twelve returned systems whose reference usage no public source documents (the unidentified systems of Section 5), for which R is undefined, sit out of this design only. The reported β is the coefficient on $D_{\text{add}} \cdot R$. Its point estimate is ordinary least squares (OLS) on the pooled 10,220 observations; every interval and p-value comes from resampling the 268 distinct source sentences with replacement ($B = 2000$ replicates) and refitting. That clustered bootstrap is the registered inference, and it carries between-source heterogeneity in intercepts and slopes alike. The registration placed this fixed part inside a mixed-effects model with random intercepts for source and for language pair; refitting that model leaves the point estimate unchanged (Appendix A). Because y is a percentile, β has a plain reading: it is the addition-minus-omission gap in mean percentile drop for reference-reading metrics minus the same gap for reference-free metrics, so $\beta = -0.069$ says the gap is 6.9 percentile points wider without a reference. Section 5.1 gives the four cell means behind it.

The reference-based anchor. H_B is a contrast, so it needs the reference-based family to be symmetric in fact. For metric m, let $b_{\text{om},m}$ be the mean raw oriented drop over its omission members and $b_{\text{add},m}$ the same over its addition members, both in the metric's own units (Appendix B lists them). The scale-free relative asymmetry is

$$A_m = \frac{|b_{\text{om},m}| - |b_{\text{add},m}|}{(|b_{\text{om},m}| + |b_{\text{add},m}|)/2}, \quad (2)$$

bounded in $[-2, 2]$, with $A_m := 0$ when both drops are exactly zero; a negative value means omission is penalized less than addition. Family-level A is the unweighted mean of A_m over that family's returned metrics, and the teeth statistic is the mean raw drop pooled over both edit directions in the family's native units, which must be positive for the family to anchor anything. Table 2 carries the decision rules: an equivalence margin of ± 0.30 tested by two one-sided tests at $\alpha = 0.05$ (equivalently, the 90% CI inside the margin, as the table states), the teeth precondition, and the refutation leg, which fires when the reference-based family's 95% CI lies entirely below -0.30 and thereby refutes the claim that the missing reference is what produces the asymmetry.

Multiplicity and orientation. Every returned metric with a declared orientation enters the confirmatory family of Table 2 mechanically, its size following from the roster before any score value is inspected, with Benjamini-Hochberg (BH) control (Benjamini and Hochberg, 1995); the harness fails closed on an undeclared name (Appendix A);

Quantity	Definition	Registered prediction	Decision rule and inference	Outcome
N1 β (Arm B, primary)	Coefficient on $D_{\text{add}} \cdot R$ in Eq. 1: the change in the addition-minus-omission penalty gap, on the within-metric percentile scale (0 to 1, so 0.01 is one percentile point), when the metric reads a reference (negative: the gap is wider without one)	$\beta < 0$	Ordinary least squares (OLS) point estimate; source-clustered bootstrap 95% CI (B=2000); one test in the confirmatory family	-0.069 (6.9 points) $[-0.081, -0.057]$, confirmed (§5.1)
N2 P_{flip} (Arm B, primary)	Share of matched {base, omission} pairs a metric scores at or above the base, per metric; registered for reference-free QE, exploratory for the other families	high for reference-free QE, near 0 for reference-based and LLM-judge metrics	Clustered bootstrap CI; descriptive	0.013 to 0.195 for reference-free QE; not near 0 elsewhere (§5.2)
Anchor $A_{\text{ref-based}}$	Family mean of the relative asymmetry A_m of Eq. 2 for the reference-based family, whose symmetry the H_B contrast needs	inside $[-0.30, 0.30]$; reference-free QE $A < 0$	Established if the 90% CI sits inside the margin and the teeth precondition holds (pooled drop 95% CI above 0); refutes the H_B mechanism if the 95% CI lies below -0.30; otherwise not established	-0.42 $[-0.49, -0.35]$: refutation leg fires; reference-free -0.51 (§5.1)
Strata gradient (secondary)	A and mean P_{flip} for reference-free QE within S1, S2, S3	A(S1) $<$ A(S2+S3); P_{flip}(S1) $>$ P_{flip}(S2+S3)	Ordering of point estimates, CIs reported; below 20 pairs, point only; never in the confirmatory family	holds as an ordering (§5.5)
Anti-circularity strata	β, A, P_{flip} on the 304 triples the pre-filter predicted would not break, and on the 189 real-MT-base triples	effect persists	Confirmation of H_B requires both	both hold (§5.5)
Arm A ρ	Mean over the 76 series of the per-series Spearman ρ between filler dose and oriented score, per metric	$\rho \geq 0$ for reference-free QE (H_A)	One-sided clustered bootstrap test of H_0: $\rho \geq 0$, in the confirmatory family	ρ from -0.37 to -1.00 for all 28 metrics: H_A refuted (§5.3)
Arm C τ_b	Kendall τ_b of correct versus corrupted member, tie rate reported, per metric	reference-based frontier catches, reference-free trails (H_C)	One-sided test of H_0: $\tau_b \leq 0$, in the confirmatory family	holds; two MQM-LLM variants invert (§5.4)
Multiplicity	One test per metric on Arm A (28 metrics) and Arm C (31) plus the single pooled Arm B β; family size fixed by the returned roster		Benjamini-Hochberg (BH) false discovery rate at $q = 0.05$, applied once	60 tests; 58 reject at $q < 0.001$

Table 2: The pre-registered plan at a glance. Every quantity, its decision rule, and its predicted sign were frozen before any score existed; the last column points to where Section 5 reports it. Arm B bootstraps resample its 268 distinct source sentences with replacement; Arm A and Arm C resample their own source sentences.

and everything outside that family, the strata and the sensitivity refits included, is secondary or exploratory and is never reported as a break.

Anti-circularity, with the numbers. A local CometKiwi pre-filter, which belongs to the same family as the audit target, labeled every row with a break prediction but never gated selection, and there was no selection step for it to influence. The packager ran with no per-direction cap and the set came in at 231,582 tokens against the 1M budget, so every item that passed the validity gates shipped. The one exception to that shipment is a deterministic triple dropped by a metric-free rule that keeps the LLM triple when both were built on the same source, applied before any label was written. The gates themselves see no metric score; they admitted 89 of the 861 complete LLM triples of Section 3.2 (cs-uk 40 of 334, en-zh 17 of 222,

en-de 11 of 69, en-cs 11 of 81, en-ru 6 of 30, en-ja 4 of 125). The label split of the shipped set is: Arm B, 4 of 308 triples predicted to break; Arm A, 0 of 76 series; Arm C, 89 of 753 pairs. H_B is reported as confirmed only if the effect also holds in the 304 Arm B triples predicted not to break and on the 189 triples whose bases are real WMT24 system outputs.

Sensitivity refits. Two refits of Eq. 1 qualify the primary interaction and can never rescue it: the registered z-score transform in place of the percentile transform, and a refit that adds edit magnitude (absolute whitespace-stripped character change of the member against its base) as a covariate, motivated by the measured 2.0x median edit-size imbalance between additions and omissions. The second was recorded in the decision log under 4 August; its first git timestamp is the 6 August return-ingestion commit, made before any score value was analysed (Appendix A).

5 Results

The organizers returned segment scores from 31 systems on 6 August 2026, scored under the task's segment-level quality-score protocol (its Task 2: one score per segment, higher is better). 28 systems cover all 2,886 segments; `cuni-v14-regression` covers three directions and the two Lexicala QE ensembles cover three of the four Arm-C-only directions (cs-de, en-ar, zh-ja). The roster is a superset: some teams submitted more than the allowed two systems, and near-duplicate submission-id variants remain because the organizers had not designated main submissions by the camera-ready deadline, so the family of 60 tests stands on the returned roster. Twelve systems carry opaque submission names whose reference usage no public source documents yet (the unidentified systems below); under the fail-closed rule they were declared with orientation only, in a pre-analysis commit, so they enter every per-metric endpoint mechanically but sit out the reference-availability axis, which is undefined for them. We write xCOMET-XL-QE and xCOMET-XXL-QE for the organizer baselines xCOMET XL and XXL Reference-free, and xCOMET-XL-ref and xCOMET-XXL-ref for their Reference-based counterparts. All ingestion checks passed: every returned segment identifier matched the private keymap, and per-system counts matched the organizers' coverage table exactly. Of the 60 confirmatory tests, 58 reject at $q < 0.001$, where rejection means the metric penalizes filler (Arm A), catches the corrupted numeral (Arm C), or shows a nonzero interaction (Arm B, two-sided); the two exceptions are the Arm C failures of the two MQM-LLM variants (Section 5.4).

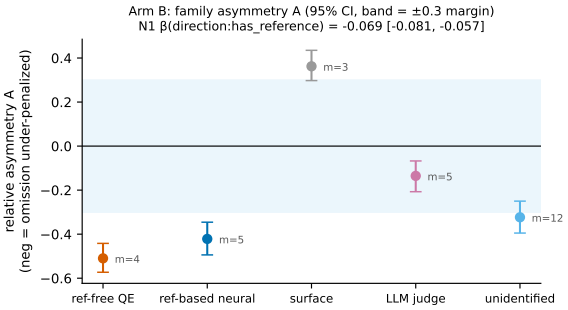

Figure 1: Family-level relative asymmetry A with clustered-bootstrap 95% CIs (values in the text); the shaded band is the pre-registered ± 0.30 equivalence margin for the reference-based anchor, and the reference-based family lies outside it.

5.1 Headline: the omission and addition asymmetry (Arm B)

The signed interaction is significant in the predicted direction: $\beta = -0.069$, source-clustered bootstrap 95% CI $[-0.081, -0.057]$, $q < 0.001$ in the confirmatory family. On its own scale (Section 4) the coefficient reads as follows. For reference-free metrics (the $R = 0$ pool, which includes the two reference-free judges) an addition sits on average at the 56.1st percentile of that metric's drops and an omission at the 43.9th, a gap of 12.2 points; for reference-reading metrics the gap is 5.2 points (52.6 against 47.4); β is the difference of the two gaps before rounding, -6.9 points ($5.24 - 12.17$). Per family, the gap is 18.3 points for reference-free QE, 13.7 for reference-based neural, 3.4 for LLM judges, and -8.6 for surface metrics, which penalize omission more. The two sensitivity refits leave the interaction intact: -0.24 standard deviations $[-0.29, -0.20]$ under the z-score transform, and $\beta = -0.073$ $[-0.084, -0.061]$ with edit magnitude as a covariate, so the 2.0x character imbalance of Section 3 does not explain it.

The family asymmetries (Figure 1) tell the larger story. Reference-free QE: $A = -0.510$ $[-0.573, -0.442]$, with per-metric A_m of -1.28 (xCOMET-XL-QE), -0.59 (xCOMET-XXL-QE), -0.22 (`cuni-v14-regression`), and

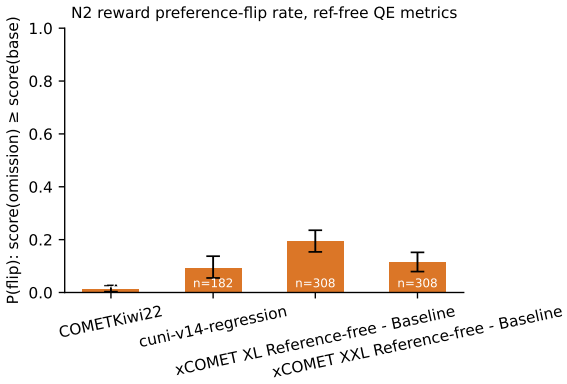

Figure 2: P_{flip} per reference-free QE metric with clustered-bootstrap 95% CIs (values in the text).

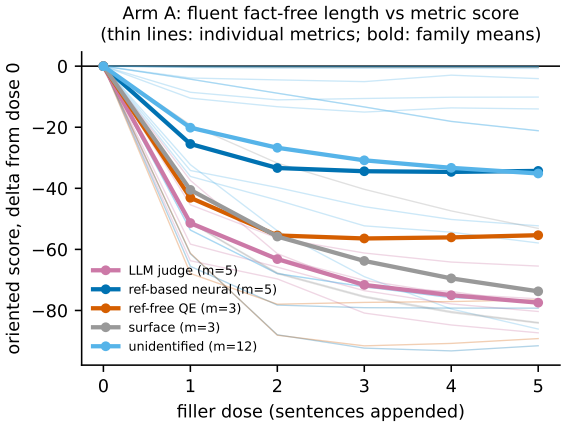

Figure 3: Oriented score against fluent fact-free filler dose (thin lines: individual metrics; bold: family means). Every family's mean falls monotonically; no returned metric rewards pure unsupported length.

$+0.05$ (COMETKiwi22). Reference-based neural: $A = -0.422$ $[-0.494, -0.346]$. LLM judges: $A = -0.135$ $[-0.207, -0.067]$. The twelve unidentified systems: $A = -0.323$ $[-0.395, -0.250]$. Surface metrics alone over-penalize omission: $A = +0.362$ $[+0.297, +0.435]$, as n-gram precision plus a brevity penalty punishes missing content twice. The anchor verdict is this paper's central honest finding. The teeth precondition holds: the pooled reference-based drop, in the family's native units and dominated by the two 0-to-100 xCOMET scales, has 95% CI $[+11.5, +13.4]$, and every member's own mean drop is positive in both edit directions (Table 4). But the reference-based 90% and 95% CIs lie entirely below the -0.30 margin, so the pre-registered refutation leg of Table 2 fires. H_B's mechanism claim is refuted: the asymmetry appears in every learned family, so the missing reference cannot be its cause. Reference access reduces it (the significant interaction) and surface overlap inverts it.

5.2 The reward preference flip (N2)

P_{flip} is the deployment-relevant number: the rate at which a reference-free metric scores the omission at or above its own base, which is exactly the contest a best-of-n reranker adjudicates. Over 308 matched pairs (182 for the partial-coverage CUNI system), Figure 2: xCOMET-XL-QE 0.195 $[0.154, 0.235]$, xCOMET-XXL-QE 0.114 $[0.079, 0.152]$, `cuni-v14-regression` 0.093 $[0.055, 0.137]$, COMETKiwi22 0.013 $[0.003, 0.027]$. Part of the first three rates is exact ties, mostly at the scale ceiling: strictly above the base, the rates are 0.143, 0.071, 0.044, and 0.013. A reranker guided by xCOMET-XL-QE rates the translation that silently dropped content at least as high as its base in one matched contest out of five, and strictly higher in one out of seven. COMETKiwi22 almost never flips and is also the family's most symmetric member (the per-metric A_m above); the deep asymmetry is carried by the xCOMET QE baselines. The failure is graded, not uniform, inside the family. The registration also expected the rate to be near 0 for reference-based and LLM-judge metrics; computed the same way for every metric (exploratory, Table 4), it is not: family means 0.088 for reference-based neural (xCOMET-XL-ref 0.208), 0.130 for LLM judges (gemba-poly 0.325), 0.074 for surface, 0.218 for the unidentified systems, against 0.104 for reference-free QE. The flip, like the asymmetry, is shared by learned metrics as a class.

5.3 Length bias (Arm A)

H_A is refuted outright, and we report the pre-registered negative as good news for the field: every returned metric penalizes fluent fact-free filler (Figure 3), with within-source Spearman ρ (the mean over series of the per-series correlation) from -0.37 (fluency2-gemini35-esa) to -1.00 (BLEU, chrF, spBLEU, YiSi-1), all $q < 0.001$ against H_0: $\rho \geq 0$. COMETKiwi22 reaches $\rho = -0.86$ $[-0.89, -0.83]$; xCOMET-XL-QE and xCOMET-XXL-QE -0.65 and -0.70. The length-gaming attack this arm operationalizes does not work on the modern roster. Reference-based curves are exploratory context on this arm (reference-verbatim bases inflate the dose-0 intercept), and H_A conclusions are scoped to en-de and

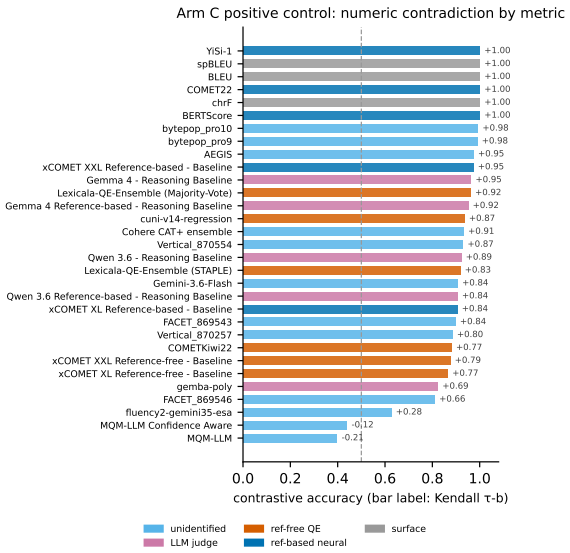

Figure 4: Numeric-contradiction contrastive accuracy, sorted; bar labels give Kendall τ_b. All families sit far above chance except the two MQM-LLM variants at the bottom, which invert the control.

en-ru as registered.

5.4 Positive control (Arm C)

The control holds: the harness has teeth (Figure 4). The reference-based neural frontier catches the single corrupted numeral (COMET22, BERTScore, and YiSi-1 at $\tau_b = +1.00$; xCOMET-XXL-ref $+0.95$), surface metrics detect the token change deterministically ($+1.00$), and the reference-free baselines trail visibly (COMETKiwi22 $+0.77$, xCOMET-XL-QE $+0.77$, xCOMET-XXL-QE $+0.79$): reference-based catches while reference-free trails, as predicted. One cell of the frozen prediction table was wrong here and we report it rather than reframe it: surface metrics were predicted to break on this arm and instead held perfectly; Table 3 in Appendix A marks it together with the three other contradicted cells. Contrastive accuracy here counts metric ties as half credit over the same pairs, ties included, on which τ_b is computed; tie rates stay at or below 0.19 and are highest for span-derived scorers. Two participant systems invert the control: MQM-LLM ($\tau_b = -0.21$, contrastive accuracy 0.397) and MQM-LLM Confidence Aware (-0.12, 0.439) score the corrupted member above the correct one more often than not, the only two confirmatory non-rejections on the board. The empirical orientation cross-check (a flag whenever a metric's mean oriented correct-minus-corrupted difference

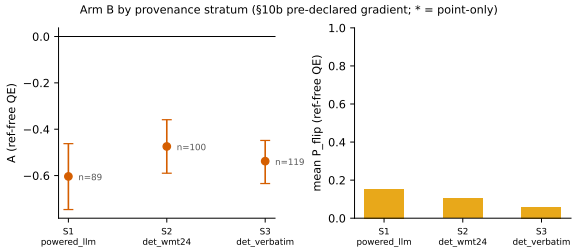

Figure 5: The pre-declared provenance-strata gradient for reference-free QE: relative asymmetry A (left) and mean P_{flip} (right). The LLM-generated stratum S1 cuts deepest on both readings, as pre-registered.

on Arm C is at or below -0.05) flags both, and their strongly negative Arm A slopes confirm the declared orientation is right, so this is a genuine system-level failure on load-bearing numerals, not a sign error in our manifest.

5.5 Stratified reading

The pre-declared gradient holds as an ordering of point estimates (Figure 5). In S1 (subtle LLM perturbations on real system outputs, 89 triples) the reference-free asymmetry deepens to $A = -0.604$ $[-0.747, -0.462]$ and mean P_{flip} rises to 0.154 $[0.107, 0.203]$, against -0.474 $[-0.590, -0.359]$ and 0.104 $[0.074, 0.135]$ in S2 and -0.538 $[-0.634, -0.449]$ and 0.058 $[0.035, 0.085]$ in S3 (S2 and S3 pooled: -0.510 and 0.079); the interaction strengthens to $\beta = -0.117$ $[-0.151, -0.082]$ inside S1. The registered prediction $A(\text{S1}) < A(\text{S2} + \text{S3})$ and $P_{\text{flip}}(\text{S1}) > P_{\text{flip}}(\text{S2} + \text{S3})$ is an ordering, and it holds on both axes; the A intervals overlap, so the gradient is a direction, not a separated effect. Both anti-circularity confirmation strata of Table 2 hold as well. In the 304 triples the local CometKiwi pre-filter predicted would not break: $\beta = -0.069$ $[-0.081, -0.057]$, $A(\text{ref-free}) = -0.503$, mean $P_{\text{flip}} = 0.094$. On the 189 triples with real WMT24 system-output bases: $\beta = -0.078$ $[-0.094, -0.061]$, $A(\text{ref-free}) = -0.487$, mean $P_{\text{flip}} = 0.129$. The measured deficit is not an artifact of pre-filter-guided construction, and it lives on real MT output; with the anchor refuted, these strata certify that the pervasive asymmetry the audit found is real wherever we can check it.

6 Human Validation Protocol

We built a blind human validation study before submission: a 270-item single-annotation sheet plus a 60-item double-annotated overlap, sampled

with a guaranteed minimum of S1 items and sealed so annotators see only an identifier, the source, and the hypothesis. The protocol targets per-direction item precision of at least 95 percent and reports Cohen's κ and Krippendorff's α on the overlap (Cohen, 1960; Krippendorff, 2018). The annotation had not been carried out by the camera-ready deadline, so this paper reports the protocol and no human precision number. That is a real gap, not a formality: the machine-side gates of Section 3.2 certify entailment relations and the absence of entities, digits, and dates, but they cannot certify that a dropped unit is one a human reader would miss, or that an appended unit reads as natural, and every per-item claim in this paper inherits that limit. The sealed sheet is available with the pipeline (Appendix A) so the study can be run independently of this paper's claims.

7 Conclusion

`QEbreak` turns a structural suspicion about reference-free QE metrics into a falsifiable, preregistered measurement: matched same-source omission and addition contrasts, carried through the reference-availability axis, down to the reward preference flip that a metric-guided reranker would actually commit. The returned scores confirmed the deficit and refuted our favorite explanation of it. Reference-free QE metrics prefer the incomplete translation often enough to matter, but so, to a smaller degree, do the reference-based metrics that were supposed to anchor the contrast. The blind spot therefore belongs to learned metrics as a class rather than to reference-freeness, and the significant interaction shows that reference access narrows it without closing it. What we would hand next year's breakers is the discipline rather than the hypothesis: the blind protocol with a private keymap, the fail-closed orientation manifest, and the equivalence anchor with its teeth precondition, which together let a wrong mechanism story die in public while the measurement that killed it stands.

Limitations

Item validity rests on machine gates. The human study of Section 6 was not run, so no per-item precision number exists; the NLI and NER gates are the operative validity evidence, with the limits stated there. Numeric pairs use Latin and Cyrillic digit contexts only; native-digit corruption for zh, ja, ar, and ko scripts was not attempted.

Coverage, stratum sizes, edit size, and dose. Arm B covers 6 of the 10 directions; cs-de, en-ar, en-ko, and zh-ja carry Arm C only. Sustained 8B-model generation on a fanless 8 GB machine limited S1 on some directions, so S1 is small on en-ja (4 triples) and en-ru (6); the registered small-stratum guard gives point estimates only to any stratum below 20 matched pairs, and any per-direction reading is exploratory. The omission and addition edits are matched on information units, not characters: additions change a median of 2.0 times as many whitespace-stripped characters as omissions remove (S1 1.4x, S2 2.0x, S3 2.5x), and 79 percent of triples add more characters than they remove; if a metric responds to raw edit magnitude this pushes A in the predicted negative direction, which is why the edit-magnitude refit is reported beside the primary interaction. Every Arm B edit carries one information unit, so no dose-interaction claim is made. Arm A is 76 series on two directions with reference-verbatim bases: for reference-free QE, the target family, base equals reference is mechanically invisible; for reference-based metrics the dose-0 intercept is inflated, so that curve is exploratory context only, and H_A conclusions are scoped to en-de and en-ru as registered.

Circularity is disclosed, not eliminated. CometKiwi labeled, never gated, Arm A and B items, and the confirmation rule requires the effect to survive in the stratum predicted not to break and on real system outputs; within-family corroboration is never treated as independent transfer. Members of a pair also share a source, group size correlates with arm, and control members are reference-verbatim. These properties are inherent to contrastive minimal-pair sets; an honest metric scores each row independently and cannot exploit them, but a deliberately adversarial scorer could. The blind protocol removes the fixable tells (labels, id tokens, positional order); the inherent ones are stated here.

Acknowledgments

We thank Eleftherios Avramidis and the WMT26 Automated Translation Quality Evaluation organizers for confirming the separate-segments design and the submission logistics, and the two anonymous reviewers, whose requests for the model specification, the anti-circularity counts, and a plan section built around one table shaped

this version.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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, Miami, Florida, USA. Association for Computational Linguistics.

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

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.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

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.

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.

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. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Yilin Zhang, Wenda Xu, Zhongtao Liu, Tetsuji Nakagawa, and Markus Freitag. 2025. Penalizing length:

Uncovering systematic bias in quality estimation metrics. *Preprint*, arXiv:2510.22028.

Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, and Daniel Deutsch. 2026. Generating difficult-to-translate texts. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 204–219, Rabat, Morocco. Association for Computational Linguistics.

A Pre-Registration Record

Frozen commits. The registration (commit `6febc4b`, 9 July 2026), a sign-gloss erratum (`b8e78f3`), the scope-and-anchor amendment (`91c19c1`), and a teeth-wording and tag-provenance clarification (`98bf78c`) all predate the 15 July upload (deadline 16 July) and the 6 August score return; the orientation manifest is frozen by content hash. The erratum corrected the verbal sign of an interpretive gloss on the interaction; it touched no endpoint, and the gloss is not used in this paper. The amendment declared the numeric equivalence bound of Eq. 2, the provenance strata, the single-dose scope of Arm B, and the two-direction scope of Arm A. The analysis code was frozen at the upload (commit `28e1706`, 15 July) apart from the post-return changes listed next; no commit lies between that upload and the return.

Post-return changes. The registration permits exactly one kind of change after scores return: adding an orientation pattern for a returned metric name, in a descendant commit made before any score value is inspected, with a decision-log entry. One commit (`f539810`, 6 August, 27 minutes after the return arrived and before any score value was analysed) made those additions: organizer baselines by their self-describing names and four participant systems from three teams (`gemba-poly`, the two Lexicala ensembles, `cuni-v14-regression`) by published descriptions; those pattern additions were needed because the harness raises on any returned metric name that matches no manifest pattern. The same commit went beyond what the registration anticipates, and we report those changes as deviations under its own rule. It declared the twelve unidentified systems as a fifth family with orientation only and changed the analysis code so that this family sits out the reference-availability design, where R is undefined, while staying in every per-metric endpoint. It added the edit-magnitude refit of Section 4 and the code for the registered fallback that reports per-metric drops in place of a family-level A when fewer than two reference-free QE metrics score Arm B; both are recorded in the decision log under 4 August with this commit as their first git timestamp, and the fallback did not fire (four scored Arm B). It added the return-ingestion code, the command-line wiring, and a plot label for the new family; plotting code changed again in the paper commit. No change touched an endpoint, a decision rule, or the confirmatory family rule. Format inspection of the return had unavoidably displayed a handful of raw rows from two files before that commit; no aggregate was computed and no declaration used score values.

Arm	Surface	Ref-based	Ref-free	LLM judge
B omission	break†	hold†	break-hard	partial
B addition	hold	partial/hold	hold/partial	hold
A verbosity	hold	partial	break†	hold
C numeric	break†	hold	break/partial	hold

Table 3: The frozen arm-by-family prediction table (registration section 2), verdict labels verbatim and the registration's one-phrase rationales omitted. "Break" means the family was predicted to score the perturbed member at or near the base (the pre-filter's operational rule in Section 3.3 is at or above); slashed cells were registered as either-or; the B omission reference-free cell was registered as the headline "break-hard". †The four cells the data contradicted: surface metrics held on Arm B omission ($A = +0.36$, Section 5.1) and on Arm C (Section 5.4); reference-free QE penalized filler on Arm A (Section 5.3); and the reference-based anchor cell fired its refutation leg (Section 5.1).

Frozen predictions. Table 3 reproduces the registered arm-by-family prediction table. Three of the four contradicted cells fail in the same direction, a predicted break that did not happen; the fourth, the Arm B reference-based cell, is different in kind: the registration built in the equivalence anchor precisely so that this cell could refute the mechanism claim, and it did.

Estimator check. The registration's mixed-effects form for Eq. 1, refitted by restricted maximum likelihood with source nested in language pair, returns $\beta = -0.069$ with Wald 95% CI $[-0.086, -0.052]$; adding random slopes for edit direction and reference availability by source gives -0.070 $[-0.086, -0.054]$. The registered clustered bootstrap, the least-squares point estimate, and both mixed-model refits agree, so nothing

Metric	Family	R	Pairs	$b_{\text{om},m}$	$b_{\text{add},m}$	A_m	Gap (pts)	Flip
COMETKiwi22	ref-free QE	0	308	+0.185	+0.177	+0.05	+0.2	0.013
cuni-v14-regression	ref-free QE	0	182	+15.9	+19.9	-0.22	+7.7	0.093
xCOMET XL Reference-free - Baseline	ref-free QE	0	308	+10.2	+46.1	-1.28	+38.1	0.195
xCOMET XXL Reference-free - Baseline	ref-free QE	0	308	+28.9	+52.9	-0.59	+22.8	0.114
BERTScore	ref-based neural	1	308	+0.070	+0.090	-0.25	+9.9	0.071
COMET22	ref-based neural	1	308	+0.116	+0.151	-0.27	+8.9	0.045
xCOMET XL Reference-based - Baseline	ref-based neural	1	308	+7.8	+40.7	-1.36	+38.7	0.208
xCOMET XXL Reference-based - Baseline	ref-based neural	1	308	+25.6	+49.6	-0.64	+23.9	0.088
YiSi-1	ref-based neural	1	308	+0.038	+0.025	+0.41	-12.7	0.029
BLEU	surface	1	308	+28.2	+25.1	+0.12	-1.3	0.117
chrF	surface	1	308	+23.3	+9.4	+0.85	-21.5	0.042
spBLEU	surface	1	308	+28.4	+25.1	+0.12	-3.1	0.062
gemba-poly	LLM judge	1	308	+13.6	+23.7	-0.54	+9.2	0.325
Gemma 4 - Reasoning Baseline	LLM judge	0	308	+38.8	+43.8	-0.12	+5.8	0.062
Gemma 4 Reference-based - Reasoning Baseline	LLM judge	1	308	+39.4	+46.0	-0.15	+8.5	0.075
Qwen 3.6 - Reasoning Baseline	LLM judge	0	308	+38.3	+35.5	+0.08	-3.5	0.101
Qwen 3.6 Reference-based - Reasoning Baseline	LLM judge	1	308	+38.7	+36.2	+0.07	-2.9	0.088
AEGIS	unidentified		308	+19.9	+22.3	-0.11	+5.7	0.023
bytepop_pro10	unidentified		308	+29.2	+39.2	-0.29	+13.7	0.019
bytepop_pro9	unidentified		308	+29.1	+39.0	-0.29	+13.6	0.019
Cohere CAT+ ensemble	unidentified		308	+29.2	+25.6	+0.13	-5.9	0.094
FACET_869543	unidentified		308	+0.149	+3.6	-1.84	+35.2	0.666
FACET_869546	unidentified		308	+0.211	+3.6	-1.78	+33.3	0.649
fluency2-gemini35-esa	unidentified		308	+7.5	+7.7	-0.03	+5.8	0.250
Gemini-3.6-Flash	unidentified		308	+32.4	+27.2	+0.18	-8.9	0.081
MQM-LLM	unidentified		308	+13.8	+13.7	+0.00	+13.3	0.292
MQM-LLM Confidence Aware	unidentified		308	+10.8	+14.0	-0.26	+10.8	0.260
Vertical_870257	unidentified		308	+0.126	+0.070	+0.56	-7.6	0.114
Vertical_870554	unidentified		308	+0.170	+0.197	-0.15	+6.4	0.146

Table 4: Per-metric Arm B quantities. R is the reference-availability indicator of Eq. 1, left blank for the unidentified systems, which sit out that design. Drops are in each metric's native units, positive when the metric penalized the edit; the gap is the mean within-metric percentile of addition drops minus that of omission drops, positive when addition is penalized more; Flip is the share of matched pairs in which the metric scores the omission at or above the base (P_{flip} for reference-free QE, exploratory elsewhere). A_m compares mean raw drops (positive when omission is penalized more) and the gap compares mean ranks, so the two agree when their signs are opposite; a same-sign pair with a gap far from zero, as for MQM-LLM (A_m near zero, gap $+13.3$), means a few large omission drops dominate the raw mean while most omissions rank low. Family A in Section 5.1 is the unweighted mean of A_m within each family.

turns on the estimator.

Availability. The build pipeline, the analysis code with its timestamped history, the labeled challenge set, the private scoring keymap, and the sealed validation sheet are available from the author on request; the organizers had not designated main submissions or a release schedule for the returned scores by the camera-ready deadline.

B Per-Metric Arm B Results

Table 4 lists, for every returned metric that scored Arm B, the mean raw oriented drops $b_{\text{om},m}$ and $b_{\text{add},m}$ of Eq. 2 in the metric's own units, the relative asymmetry A_m, the addition-minus-omission gap in within-metric percentile points that Eq. 1 averages, and the flip rate (the share of matched pairs scoring the omission at or above the base, which is P_{flip} for the reference-free QE family and exploratory elsewhere). The two Lexicala ensembles did not score any Arm B direction.