The Blindness of Document-Level Translation Evaluation

Ahrii Kim[1] **Vilém Zouhar**[4] **Chanjun Park**[2] **Seong-heum Kim**[1,3] *
[1]AI-Bio Convergence Research Inst. [2]School of Software
[3]Dept. of Intelligent Semiconductors [4]ETH Zurich
Soongsil University
{ahriikim,chanjun.park,seongheum}@ssu.ac.kr vzouhar@ethz.ch

Abstract

Document-level machine translation (MT) evaluation extends segment-level protocols by presenting full documents to annotators, on the assumption that such presentation elicits document-level judgments. We test this assumption with a counterfactual condition (**MIX**) in which each document combines segments drawn from different systems, preserving document-level presentation while breaking cross-segment consistency. Across 18,420 expert English→Korean annotations and 14 automatic metrics, scores, system rankings, and error annotations are statistically equivalent between coherent and incoherent documents. Perception does not explain this: shown matched passages, raters identify the coherent one as the work of a single translator in 87.3% of trials. Document presentation does change how annotators work, but that change does not reach the recorded output. What is blind is the protocol, not the annotator. The concern is not that scores fall short, but that the resources invested in document-level systems, metrics, and annotation may not be measuring what they are intended to measure.

1 Introduction

Document-level machine translation (MT) has been described as the missing piece for reaching human-level translation performance (Kocmi et al., 2025). Early claims of human parity, on closer inspection, reflected the mismatch between segment-level and document-level translation (Läubli et al., 2018; Popel et al., 2020; Toral, 2020). Yet a series of puzzling patterns has accumulated in document-level MT evaluation: non-robust system rankings (Freitag et al., 2024; Kocmi et al., 2024a, 2025), human scores clustered near the top of the scale (Kocmi et al., 2024a), surface metrics like chrF (Popović, 2015) performing comparably to large language model (LLM)-based ones (Lavie et al., 2025), and literary text emerging as the easiest domain (Kocmi et al., 2025). These anomalies persist in part because the validity of current document-level evaluation protocols has not been thoroughly confirmed.

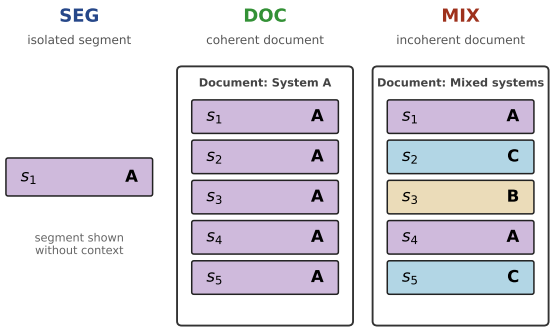

Figure 1: The three evaluation conditions. **SEG** presents a segment in isolation, without surrounding context. **DOC** presents a coherent document produced by a single system (here, System A). **MIX** presents a document of the same form but assembled from segments of different systems (A, B, C), preserving document-level layout while disrupting cross-segment coherence.

Cognitive science offers a useful lens. Human evaluators can fail to notice quality issues through *content-driven blindness*, where familiar or expected content reduces scrutiny (Nickerson, 1998), or through *presentation-driven blindness*, where attention is shaped by how information is structured or displayed (Mack and Rock, 1998; Simons and Chabris, 1999; Treisman and Gelade, 1980). In MT evaluation, both raise the possibility that document-level cues go unregistered. The distinction we will need throughout is between document-level *presentation*, which current protocols do provide, and document-level *judgment*, which they assume follows from it. A cue can be perceived and still go unregistered if the instrument offers nowhere to put

*Corresponding author.

https://huggingface.co/datasets/trotacodigos/esa-counterfactual-enko

https://github.com/trotacodigos/esa-counterfactual

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 12433–12454
October 24-29, 2026 ©2026 Association for Computational Linguistics

it. Our results place the failure at that second step: what is blind is the protocol, not the annotator.

To probe this possibility experimentally, we introduce a controlled counterfactual condition (MIX) in which each document combines segments (sentences or paragraphs) from multiple MT systems, preserving document-level presentation while intentionally removing discourse coherence (Figure 1). We compare evaluation behavior and quality on MIX against coherent documents (DOC) and isolated segments (SEG) under both human annotation with Error Span Annotation (ESA; Kocmi et al., 2024b) and 14 automatic metrics. We invest in depth over breadth, collecting 18,420 expert annotations on English→Korean under a within-annotator paired design.Korean's discourse-sensitive features offer a stress test for ESA, and establishing equivalence requires statistical power that thinner multi-language designs cannot achieve at comparable cost. A blind positive control confirms that the manipulation lands: shown a DOC and a MIX passage matched on segment quality, raters identify DOC as the work of a single consistent translator in 87.3% of trials. We find that:

1. Human ESA annotation yields statistically equivalent scores, near-identical system rankings, and largely indistinguishable error annotations between coherent (DOC) and incoherent (MIX) documents.
2. All 14 automatic metrics tested produce the same system rankings on DOC and MIX. Only human evaluation shows slight divergence.
3. The equivalence is not explained by annotators ignoring context. Score recoverability, agreement, and time-on-task all indicate that document presentation *does* change annotator behavior, and ESA simply provides no channel in which that change can register.

As context-aware modeling matures, the construct-validity of document-level evaluation becomes more urgent. Our results suggest that, under current evaluation protocols, much of what is reported as document-level progress may reflect segment-level signal aggregated at the document level. The concern is not that a measure has been corrupted by being targeted, but that the measure and the target are misaligned (Goodhart, 1984). We take this as a reason to revisit discourse-level evaluation before it further misinforms the efforts built on top of it.

2 Related Work

2.1 Human Evaluation of Document MT

Human evaluation has evolved through several protocols, all sharing a segment-level core. Direct Assessment (DA; Graham et al., 2013) aggregates segment-level scores into document-level averages (Bojar et al., 2017). Early holistic document-level scoring suffered from low inter-annotator agreement (IAA; Läubli et al., 2018; Castilho, 2020), and the field subsequently converged on segment-level annotation with document context (Kocmi et al., 2022). Within this paradigm, Multidimensional Quality Metrics (MQM; Freitag et al., 2021), ESA (Kocmi et al., 2024b), and RATE (Popov et al., 2025) have refined error taxonomies and annotation interfaces, while recent work has expanded the evaluation criteria to include document-level error categories (Kim, 2025a; Song et al., 2025). Still, the existing approaches remain segment-centric: document context is provided as a cue for segment scoring, with no direct evidence that the resulting annotations capture document-level quality.

2.2 Automatic Evaluation of Document MT

Context-Extended Metrics. One line of work extends existing metrics with broader context. d-BLEU (Liu et al., 2020) and doc-COMET (Vernikos et al., 2022) operate on full documents, while training-free alternatives prepend preceding context to pretrained metrics (Vernikos et al., 2022) or aggregate adjacent sentences via sliding windows (SLIDE; Raunak et al., 2023). SEGALE (Wang et al., 2025) scales this approach to book-length documents through adaptive sentence alignment. Local windows, however, may not capture phenomena requiring global coherence, and for full-document inputs it remains unclear whether the metrics encode document-level information or merely exploit extended surface context.

Discourse-Targeted Metrics. A second line targets specific phenomena: pronoun resolution (AUTOPRF, Hardmeier and Federico, 2010; APT, Miculicich Werlen and Popescu-Belis, 2017; PROTEST, Guillou and Hardmeier, 2018), lexical choice (ATEC, Wong and Kit, 2010), and multi-phenomenon evaluation covering tense, named entities, and terminology (Sun et al., 2022; Jiang et al., 2022; Wang et al., 2023). These metrics offer fine-grained coverage but rely on surface-level string

matching, have been validated only on their target languages, and do not address cross-sentential phenomena such as narrative coherence. MUDA (Fernandes et al., 2023) instead releases taggers that locate context-dependent phenomena in arbitrary datasets, offering a route to quantifying per phenomenon how much inconsistency our counterfactual introduces.

LLM-as-a-judge. LLM-as-judge approaches have recently been applied to discourse-level evaluation, prompting LLMs to score document-level phenomena such as coherence (Kim, 2025b; Sun et al., 2025) or specific properties such as coreference and formality (Mohammed et al., 2026). These methods rely on loosely specified criteria that have produced low IAA and have been reported to perform inconsistently on discourse-level phenomena (Xiao et al., 2011; Sun et al., 2025; Kim, 2025a).

Perturbation has been used to probe context sensitivity on the modeling side. Mohammed and Niculae (2025) perturb and randomize the document context supplied to nine LLMs and two encoder–decoder baselines, and find that gains in document-level translation quality do not carry over to pronoun translation. The logic is the one we adopt: a system that genuinely uses context should react when that context is corrupted. We apply it one level up, to the instrument rather than the model. Across all lines of work, an assumption has gone untested: that providing document-level input, whether as annotator context or as extended metric input, is sufficient for evaluation to operate at the document level. We test this assumption experimentally.

3 Methods

The ESA protocol. Given a document d translated by system $s \in S$ and split into n segments, ESA elicits one score per segment while presenting the full document as context (**DOC**). Let $e_{s,k}$ denote the score assigned to segment k under system s. The document-level score is the average of segment scores:

$$E_s(d) = \frac{1}{n}\sum_{k=1}^{n} e_{s,k} \tag{1}$$

The design intent of ESA is that document-level presentation allows annotators to register information not available from any segment in isolation (**SEG**). To make this intent explicit, we decompose each segment score as

$$e_{s,k} = e^{\text{seg}}_{s,k} + e^{\text{ctx}}_{s,k}, \tag{2}$$

where $e^{\text{seg}}_{s,k}$ is the score the segment would receive if judged in isolation, and $e^{\text{ctx}}_{s,k}$ is the document-level cues induced by access to cross-segment context. ESA thereby assumes that document-level effects are captured through $e^{\text{ctx}}_{s,k}$, provided that the full document is visible and judgments are recorded at the segment level.

The counterfactual ESA protocol. We introduce a counterfactual condition (**MIX**) that preserves the interface constraints of ESA, namely document-level display with per-segment judgments, while removing genuine document-level coherence by construction. A mixed document d^{mix} is assembled from segments of multiple systems, retaining each segment's within-document position but disrupting cross-segment coherence. Annotators provide segment scores $e^{\text{mix}}_{s,k}$ under the standard ESA interface. The counterfactual document score is defined analogously:

$$E^{\text{mix}}_s(d) = \frac{1}{n}\sum_{k=1}^{n} e^{\text{mix}}_{s,k} \tag{3}$$

Using the same additive decomposition,

$$e^{\text{mix}}_{s,k} = e^{\text{seg}}_{s,k} + e^{\text{ctx,mix}}_{s,k} \tag{4}$$

where $e^{\text{ctx,mix}}_{s,k}$ captures any context-driven cues induced by the mixed presentation. If ESA registers cross-segment coherence as intended, the presence of incoherent context under **MIX** should drive $e^{\text{ctx,mix}}_{s,k}$ in a direction distinguishable from $e^{\text{ctx}}_{s,k}$, and we know that $e^{\text{ctx,mix}}_{s,k} \neq 0$.

Context cues. ESA produces three outputs per segment that can carry $e^{\text{ctx,mix}}_{s,k}$: ① a numeric score on a 100-point scale; ② error span annotations, in which annotators highlight problematic spans; and ③ a severity label (Major/Minor) on each span. If $e^{\text{ctx,mix}}_{s,k}$ differs from $e^{\text{ctx}}_{s,k}$, we expect **MIX** to depart from **DOC** along these signals. Scores should be lower under **MIX**, errors should be more frequent and more severe, and highlighted spans would concentrate at segment boundaries, where transitions between adjacent sentences are most visible.

4 Experiment

4.1 Hypotheses

We formalize three predictions entailed by ESA's design, if it registers cross-segment coherence as intended, for the contrast between **DOC** and **MIX** using three signals introduced above (①, ②, and ③). For condition c, let μ_c denote the mean segment-level score, r_c the induced system ranking, and η_c the mean error count per annotation. $\tau_{a,b} = \tau(r_a, r_b)$ denotes Kendall's τ between rankings under conditions a and b. We pre-specify equivalence margins ϵ and ϵ_{err} for scores and error counts, and a lower bound $\tau_{\min}$ for ranking agreement. Subscripts DOC and MIX refer to **DOC** and **MIX**.

H1: Score Equivalence

$$H_0 : |\mu_{\text{DOC}} - \mu_{\text{MIX}}| \geq \epsilon$$
$$H_1 : |\mu_{\text{DOC}} - \mu_{\text{MIX}}| < \epsilon$$

H2: Ranking Agreement

$$H_0 : \tau_{\text{DOC,MIX}} < \tau_{\min}$$
$$H_1 : \tau_{\text{DOC,MIX}} \geq \tau_{\min}$$

H3: Error Equivalence

$$H_0 : |\eta_{\text{MIX}} - \eta_{\text{DOC}}| \geq \epsilon_{\text{err}}$$
$$H_1 : |\eta_{\text{MIX}} - \eta_{\text{DOC}}| < \epsilon_{\text{err}}$$

Our primary comparison is between **DOC** and **MIX**. We additionally compare **SEG** with **DOC** to check whether the document-level presentation format affects evaluation outcomes independently of coherence.

4.2 Dataset

We base our study on the WMT 2025 English→Korean test set (Kocmi et al., 2025) in *news*, *social*, and *literary* domains. We exclude the *speech* domain, whose single-utterance composition is unsuitable for document-level analysis. Korean is known to pose substantial challenges for document-level translation, with discourse-sensitive phenomena including pro-drop, honorifics, and cross-segment entity consistency (Lee et al., 2025; Park and Padó, 2024; Choi et al., 2018). After filtering and length balancing (details in Appendix A.1), we obtain 3,070 unique translations from 9 MT systems and one human reference (👤REF), across 23 documents. Each segment is evaluated under all three conditions by two annotators, yielding 18,420 expert judgments, matching the scale of a full WMT annotation campaign of a single language pair.

4.3 Manipulation checks

The counterfactual is only informative if **MIX** really is the incoherent condition it is meant to be. We ran two separate studies, on different comparison units, that answer different questions. Details are discussed in Appendix A.2.

Do the pooled systems differ? The first study compares *system pairs*. For the three pairs with the highest MQM rank overlap, raters saw a source segment alongside two anonymized translations and judged which was better, or whether they were indistinguishable (486 judgments, three raters). A quality difference was perceived in 85–94% of judgments, even for these maximally overlapping pairs. The segments pooled into **MIX** are therefore genuinely different translations, not near-identical ones.

Does the assembled document read as inconsistent? Difference between systems is not inconsistency within a document, and only the second property is what **MIX** is meant to create. The second study therefore compares *documents*. We sampled 102 matched passage pairs, each showing the same six contiguous source segments twice, once as **DOC** and once as **MIX**. Three raters, blind to the manipulation, chose which passage reads as the work of a single consistent translator. **DOC** was chosen in 87.3% of trials (95% CI $[79.4, 92.4]$; Krippendorff's $\alpha = 0.73$; binomial $p < 0.001$). Readers asked about consistency therefore separate the two conditions reliably.

Two consequences follow for how the results below should be read. A weak manipulation is ruled out as an explanation of any **DOC**–**MIX** equivalence, though other explanations are not. And a null becomes interpretable: had annotators been unable to perceive the disruption, this control would have returned chance performance. It did not. The disruption is available to be perceived, so the question is whether ESA records it.

4.4 Human Evaluation

We use the ESA annotation protocol (Kocmi et al., 2024b) on the Pearmut annotation platform (Zouhar and Kocmi, 2026). Each rater annotates 706 tasks drawn randomly from the three conditions, blinded to condition assignment. We recruited 10 professional En→Ko translators, all

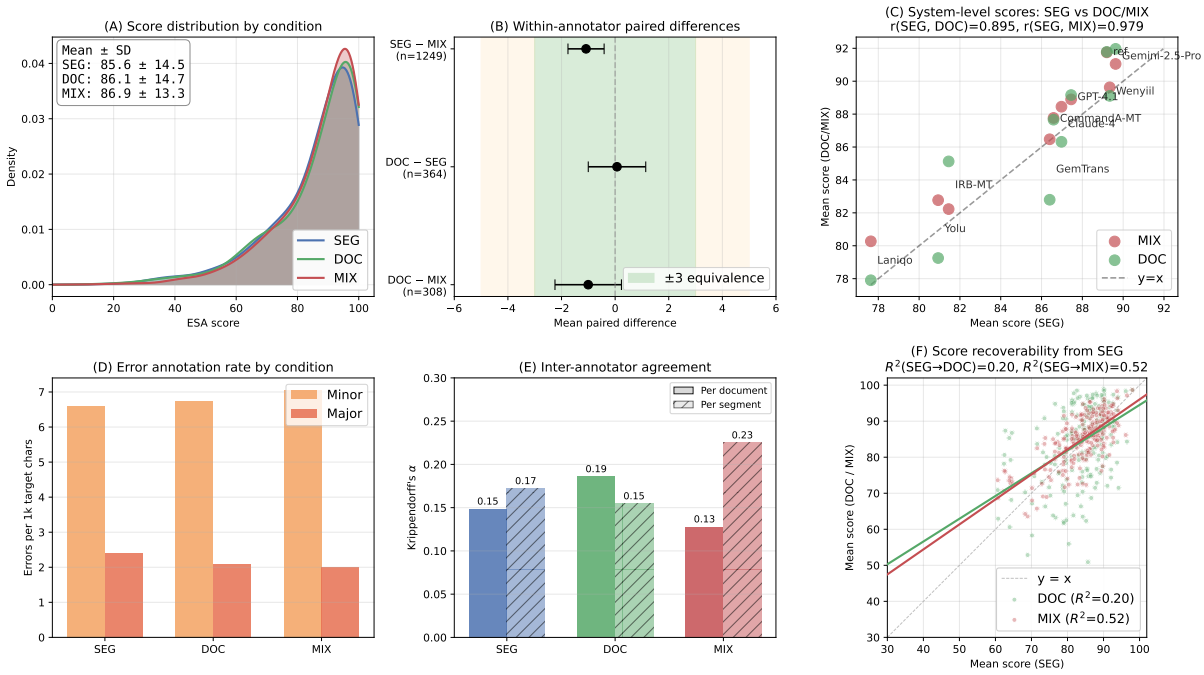

Figure 2: Overview of human ESA evaluation results across three conditions (**SEG**, **DOC**, **MIX**). **(A)** Score distributions (KDE) are nearly identical across conditions; mean and SD shown in inset. **(B)** Within-annotator paired differences fall within the ± 3 equivalence band for all three comparisons (H1); each is signed as labelled, e.g. **DOC** − **MIX**. **(C)** System-level mean scores: **SEG** vs. document-level conditions; points near $y = x$ indicate that doc-level rankings closely follow segment-level rankings (H2). **(D)** Error annotation rate per 1,000 target characters by severity (H3). **(E)** Krippendorff's α computed per document and per segment. The two units yield opposite orderings: **DOC** agrees most per document but least per segment. **(F)** Score recoverability: **SEG** scores predict 52% of the variance in **MIX** scores but only 20% of **DOC** scores ($n = 230$ matched pairs per regression).

native speakers of Korean. Further details are reported in Appendix A.3.

4.5 Automatic Metrics

Segment-level. chrF (Popović, 2015); the learned metrics XCOMET-XXL (Guerreiro et al., 2024) and METRICX-24-XXL (Juraska et al., 2024), together with their Quality Estimation (QE) variants COMETKIWI-XXL (Rei et al., 2023) and METRICX-24-QE-XXL; and the LLM-as-judge metric GEMBA V2 (Junczys-Dowmunt, 2025).

Document-level. As most discourse-targeted metrics do not support Korean, we use d-BLEU (Liu et al., 2020), doc-COMET (Vernikos et al., 2022), SLIDE (Raunak et al., 2023), and the LLM-as-judge metric FALCON (Kim, 2025b). We also include document-level variants of XCOMET-XXL and METRICX-24-XXL (and their QE counterparts), denoted with a -doc suffix.

4.6 Testing Procedure

Hypothesis testing. We test each hypothesis with a single targeted procedure. For H1, we apply the Two One-Sided Tests (TOST) procedure (Schuirmann, 1987; Lakens, 2017) to assess whether **DOC** and **MIX** scores are statistically equivalent within $\epsilon = 3$. For H2, we compute Kendall's τ (Kendall, 1938) between r_{DOC} and r_{MIX} on system-level mean scores with bootstrap confidence intervals, and compare the lower bound against $\tau_{\min} = 0.667$, the conventional two-thirds threshold for acceptable ranking agreement. Because raw rank correlations can overstate differences between statistically indistinguishable systems, we additionally cluster systems via pairwise bootstrap tests following WMT methodology (Kocmi et al., 2025). For H3, we apply paired TOST on per-segment error counts with $\epsilon_{\text{err}} = 0.5$ errors, roughly 20% of the baseline rate of 2.3 errors per annotation. Full methodological details appear in Appendix B.

Behavioral measures. Beyond hypothesis testing, we measure IAA (Krippendorff's α) and per-segment annotation time as complementary evidence on annotator's cognitive efforts across conditions.

5 Results

Figure 2 summarizes our human evaluation results. The kernel density estimates (KDE) in panel (A) show that score distributions are nearly identical across all three conditions. This is the central pattern of the section: **DOC** and **MIX** behave equivalently under ESA evaluation, both for human annotators and for automatic metrics. The following subsections present each result by hypothesis (Sections 5.1 to 5.3), with the mechanism analyzed in Section 5.4. Behavioral analyses and domain-wise breakdowns appear in Section 5.5 and Section 5.6. H1 and H2 apply to both human ESA scores and automatic metric scores, and H3 applies only to human annotations, since most metrics tested do not produce error annotations.

5.1 [H1] Scores are insensitive to discourse

Statistically equivalent scores. Table 1 reports paired TOST results across human ESA scores and 14 automatic metrics. For human ESA, mean scores are nearly identical across conditions (**SEG**: 85.56, **DOC**: 86.11, **MIX**: 86.93), and within-pair differences are all within 1.1 points (Figure 2B). TOST rejects the null at $\epsilon = 3$ for every pairwise comparison ($p < 0.001$), establishing equivalence within ± 3 points. We note that **MIX** scores marginally exceed those of **DOC**, the opposite of the predicted direction, although the two conditions remain statistically indistinguishable. The same equivalence holds across all automatic metrics: mean differences between **DOC** and **MIX** are small relative to each metric's scale, and TOST rejects the null at the per-scale equivalence margin for every metric ($p < 0.001$).

Document-level metrics react in the wrong direction. All 14 metrics support equivalence between **DOC** and **MIX** (Table 1). Among them, the metrics with the largest raw differences are precisely the document-level ones: METRICX-24-QE-XXL-doc ($d = -0.69$), METRICX-24-XXL-doc ($d = -0.63$), and d-BLEU ($d = -0.28$). This negative sign which is also observed in human scores, however, means that **MIX** receives marginally higher scores than **DOC**, the opposite of the direction of genuine coherence sensitivity (H1). This reversal is more pronounced in document-level metrics than in segment-level ones. The present data do not let us identify why the reversal arises, but they do establish what ESA's outputs were not tracking. Neither the human scores nor any metric score moved with cross-segment coherence.

Evaluator / Metric	n	ε	Mean diff	90% CI
Human ESA: pairwise paired TOST				
DOC vs **MIX**	308	3.00	−1.00	[−2.05, +0.04]
DOC vs **SEG**	364	3.00	+0.07	[−0.83, +0.97]
SEG vs **MIX**	1249	3.00	−1.09	[−1.65, −0.52]
*Segment-level metrics: **DOC** vs **MIX***				
chrF	230	3.00	+0.0187	[−0.515, +0.552]
METRICX-24-XXL	230	0.50	+0.0037	[−0.113, +0.121]
METRICX-24-QE-XXL	230	0.50	−0.0009	[−0.105, +0.103]
XCOMET-XXL	230	0.05	+0.0002	[−0.016, +0.016]
COMETKIWI-XXL	230	0.05	+0.0001	[−0.006, +0.007]
GEMBA V2	230	3.00	−0.1538	[−0.892, +0.584]
*Document-level metrics: **DOC** vs **MIX***				
d-BLEU	230	3.00	−0.2795	[−0.837, +0.278]
doc-COMET	230	0.05	−0.0008	[−0.003, +0.001]
SLIDE	230	0.05	+0.0016	[−0.007, +0.010]
METRICX-24-XXL-doc	230	3.00	−0.6346	[−1.287, +0.018]
METRICX-24-QE-XXL-doc	230	3.00	−0.6914	[−1.350, −0.033]
XCOMET-XXL-doc	230	0.05	−0.0119	[−0.019, −0.004]
COMETKIWI-XXL-doc	230	3.00	+0.0104	[−0.002, +0.023]
FALCON	230	3.00	+0.0421	[−0.687, +0.771]

Table 1: **Paired TOST for H1: score equivalence across evaluators.** Human ESA reports all three pairwise comparisons; metrics report only **DOC** vs. **MIX**. TOST rejects the null at $p < 0.001$ for all rows, establishing equivalence within $\pm\varepsilon$ ($\varepsilon \in \{0.05, 0.5, 3.0\}$ depending on scale). The shaded row indicates a known-failure case for context truncation. The metrics with wrong direction are colored in red.

5.2 [H2] System rankings collapse

Rankings converge across conditions. **DOC**-based rankings are nearly indistinguishable from **MIX**-based rankings (Table 2). Human ESA yields $\tau = 0.778$, and every automatic metric yields $\tau = 1.00$.[1] The same holds for **SEG**: τ is 0.895 between **SEG** and **DOC**, and 0.979 between **SEG** and **MIX** (Figure 2C). Neither document-level condition produces a system ordering distinct from segment-level evaluation.

Statistical clusters coincide. For human ESA, **SEG** produces 4 clusters, while **DOC** and **MIX** produce 6 and 5. The cluster counts differ, but the compositions under **DOC** and **MIX** are nearly the same. The top cluster (GEMINI-2.5-PRO, 👤REF) and the bottom-ranked system (LANIQO) are shared, and 8 of 10 systems sit at identical or adjacent ranks across the two conditions. Automatic metrics show the same pattern (Table 2). Cluster counts vary minimally between **DOC** and **MIX** across all metrics. The extra granularity that **DOC** and **MIX** provide over **SEG** is therefore not driven by cross-segment coherence. Document-level presentation produces

[1] The bootstrap CI for human ESA is $[0.539, 0.911]$, whose lower bound does not clear $\tau_{\min}$. See Appendix B.2, where the clustering result is the more informative comparison.

System	Human			chrF		xCOMET-XXL		CometKiwi-XXL		MetricX-24-XXL *		MetricX-24-QE-XXL *		GEMBA v2	
	SEG	DOC	MIX	DOC	MIX	DOC	MIX	DOC	MIX	DOC	MIX	DOC	MIX	DOC	MIX
Gemini-2.5-Pro	89.63 (1)	91.97 (1)	91.05 (2)	33.18 (1)	33.21 (1)	.791 (1)	.792 (1)	.845 (3)	.845 (**2**)	3.83 (1)	3.81 (1)	3.72 (1)	3.72 (1)	87.95 (1)	87.91 (1)
👤REF	89.20 (1)	91.78 (1)	91.75 (1)	—	—	—	—	—	—	—	—	—	—	—	—
Wenyiil	89.34 (1)	89.11 (2)	89.63 (2)	30.72 (3)	30.79 (3)	.790 (1)	.787 (1)	.852 (2)	.851 (2)	3.97 (1)	3.97 (1)	3.64 (1)	3.66 (1)	86.74 (2)	86.71 (2)
GPT-4.1	87.45 (2)	89.16 (2)	88.90 (2)	30.35 (3)	30.25 (3)	.782 (1)	.782 (1)	.837 (3)	.836 (3)	4.30 (2)	4.27 (2)	3.98 (3)	3.97 (3)	86.18 (2)	86.22 (2)
Claude-4	86.98 (2)	86.32 (3)	88.45 (**2**)	32.21 (2)	32.29 (2)	.735 (3)	.734 (3)	.842 (3)	.842 (3)	4.39 (2)	4.37 (2)	4.09 (3)	4.10 (3)	85.43 (3)	85.39 (3)
CommandA-MT	86.59 (2)	87.66 (3)	87.77 (3)	29.14 (4)	29.11 (4)	.770 (2)	.772 (2)	.861 (1)	.862 (1)	4.26 (2)	4.27 (2)	3.81 (2)	3.80 (2)	85.07 (3)	85.12 (3)
GemTrans	86.40 (2)	82.80 (5)	86.47 (**3**)	28.67 (4)	28.65 (4)	.790 (1)	.791 (1)	.830 (4)	.831 (4)	3.86 (1)	3.87 (1)	3.60 (1)	3.60 (1)	83.21 (4)	83.26 (4)
IRB-MT	81.45 (3)	85.13 (4)	82.23 (4)	25.95 (6)	25.97 (6)	.716 (4)	.717 (**3**)	.806 (5)	.806 (5)	4.83 (4)	4.82 (4)	4.44 (4)	4.45 (**5**)	81.34 (5)	81.29 (5)
YOLU	80.93 (3)	79.25 (6)	82.77 (**4**)	27.64 (5)	27.58 (5)	.739 (3)	.738 (3)	.848 (2)	.848 (2)	4.50 (3)	4.50 (3)	3.88 (2)	3.88 (2)	82.86 (4)	82.90 (4)
Laniqo	77.63 (4)	77.90 (6)	80.27 (**5**)	23.36 (7)	23.41 (7)	.719 (3)	.721 (3)	.833 (4)	.833 (4)	4.86 (4)	4.84 (4)	4.12 (3)	4.12 (**4**)	79.52 (5)	79.55 (5)
# clusters	4 / 6 / 5			7 / 7		4 / 3		5 / 5		4 / 4		4 / 5		5 / 5	
τ (DOC, MIX)	0.78			1.00		1.00		1.00		1.00		1.00		1.00	

Table 2: **System rankings under each evaluator for H2.** Cell colors indicate cluster rank (darker green = higher), derived from pairwise bootstrap tests ($B = 1{,}000$, $\alpha = 0.05$); systems within a cluster are not significantly distinguishable. Bold values mark cluster shifts between **DOC** and **MIX**. Bottom rows report the number of clusters per condition and Kendall's τ between **DOC** and **MIX** system rankings. *(lower = better).

(a) Per-annotation error counts

Condition	n	Total	Major
SEG	5,920	2.36	0.63
DOC	5,920	2.31	0.55
MIX	5,920	2.37	0.53

(b) Paired TOST: total errors ($\varepsilon_{\text{err}} = 0.5$)

Comparison	n pairs	Mean diff (90% CI)	TOST p
DOC vs **MIX**	1,157	-0.15 $[-0.25, -0.05]$	<0.001
DOC vs **SEG**	1,184	$+0.02$ $[-0.06, +0.11]$	<0.001
SEG vs **MIX**	1,185	-0.08 $[-0.18, +0.01]$	<0.001

Table 3: **Error counts and paired TOST results for H3.** (a) Mean error counts per annotation across conditions. (b) Pairwise paired TOST on total errors with $\varepsilon_{\text{err}} = 0.5$. The highlighted row marks H3's primary comparison. TOST rejects the null at $p < 0.001$ for all comparisons, establishing equivalence within ± 0.5 errors per annotation.

finer system distinctions whether the document is coherent or not.

5.3 [H3] No additional errors are marked

Error counts are equivalent. Table 3 reports paired TOST results for error counts. Mean error counts per annotation are nearly identical across conditions (**SEG**: 2.36, **DOC**: 2.31, **MIX**: 2.37), with within-pair differences of at most 0.15. TOST rejects the null at $\epsilon_{\text{err}} = 0.5$ errors per annotation ($p < 0.001$) for every comparison, showing equivalence within roughly 20% of the baseline error rate. For Major errors, the direction even reverses: **MIX** shows fewer Major errors than **DOC** (0.53 vs. 0.55; Figure 2D), opposite to the predicted direction.

Error-free rates coincide. The proportion of error-free annotations is nearly identical between **DOC** (15.5%) and **MIX** (15.7%), with **SEG** slightly lower at 12.8%. Documents in **MIX** are marked as error-free as often as coherent ones, despite their deliberate incoherence.

Error spans are not boundary-localized. Signal ③ predicts that spans should concentrate where consecutive segments meet, since that is where **MIX**'s discontinuities occur. They do not. Counting a span as boundary-localized when it falls within the first or last 20% of its segment, the rate is 50.3% under **SEG**, 50.1% under **DOC** and 50.0% under **MIX** ($n = 49{,}749$ spans; **MIX** − **DOC** $= -0.1$ points, 95% CI $[-1.2, +1.0]$). The pattern is unchanged when zero-width omission markers are excluded, when the analysis is restricted to document-interior segments, and when it is restricted to *Major* spans (Appendix B.3). Signal ③ therefore fails alongside H1 to H3: whatever annotators noticed at the seams, they did not mark it where the seams are.

5.4 Mechanism: MIX is evaluated in isolation

The preceding hypotheses show that ESA fails to distinguish **DOC** from **MIX**. Two analyses indicate why: under **MIX**, both human annotators and automatic metrics fall back to segment-level evaluation. We establish this through score recoverability for human annotators and through ranking agreement for automatic metrics.

Score recoverability. We fit linear regressions predicting each document-level score from the corresponding **SEG** score, matched by (`doc_id`, `system`) ($n = 230$ matched pairs per regression; Figure 2F). **SEG** scores explain 52% of the variance in **MIX** scores ($R^2 = 0.52$, $r = 0.72$), but only 20% of **DOC** scores ($R^2 = 0.20$, $r = 0.45$). The higher R^2 for **MIX** indicates that, even with the full document on screen, annotators fall back to segment-level judgments once inter-sentence coherence is broken. When document-level cues become

uninformative, each segment is rated much as it would be in isolation.

Metric-human agreement peaks under MIX. Figure 3 shows the same pattern in human-metric system correlations: every metric of every category, from surface to learned to LLM-as-judge, lies above $y = x$. The peak occurs under the condition in which humans themselves fall back to segment-level judgments, suggesting that current metrics align with humans through a shared segment-level fallback rather than through shared sensitivity to document-level quality. The effect is largest for chrF, the most surface-level metric in the set, and smaller for learned QE metrics, which deviate least from their DOC behavior. Per-metric details appear in Appendix C.3.

5.5 Behavioral analysis

IAA. Figure 2E reports Krippendorff's α at two units of analysis, with an inverted pattern between them. Computed per document, DOC leads ($\alpha = 0.19$, vs. 0.15 for SEG and 0.13 for MIX). Computed per segment, the order reverses, with MIX leading (0.23 vs. 0.17 for SEG and 0.15 for DOC). The per-document result for MIX is constrained by construction, since MIX documents differ across annotators by design (Appendix C.2).

This bears on an alternative reading of Section 5.4, on which the extra variance in DOC scores is annotator noise rather than context use. Noise would depress agreement at *both* units. Instead DOC agrees most per document and least per segment: annotators diverge on individual segments yet converge on overall impressions, which is the signature of context processed and expressed idiosyncratically rather than of random noise. IAA does not directly measure context use, so we treat this as corroborating rather than decisive.

Annotation time. Annotators spend less time per segment under document-level conditions than under SEG (61s for DOC and 56s for MIX, against 76s for SEG; Table 4). The decrease suggests that document context accelerates processing whether or not the context is coherent. This fits the segment-level fallback observed in Section 5.4. When annotators evaluate at the segment level, surrounding context, coherent or not, acts as a processing shortcut rather than as new information to evaluate.

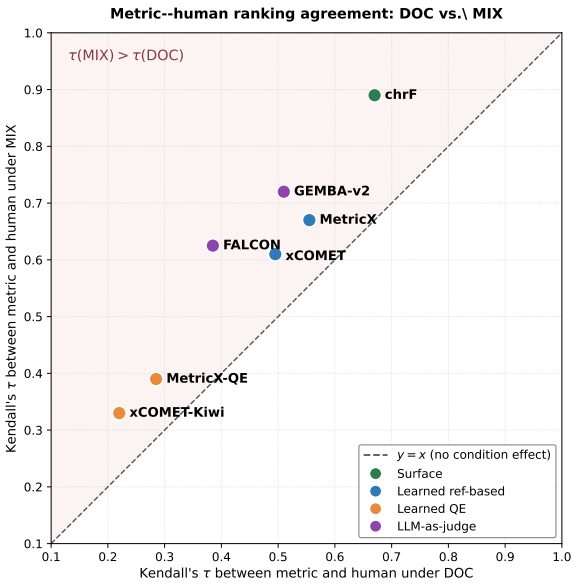

Figure 3: **Metric↔human ranking agreement under DOC and MIX.** Each point shows Kendall's τ between a metric's system ranking and the human ranking, under DOC (x-axis) and MIX (y-axis). All metrics lie above $y = x$ (shaded region), meaning that every metric agrees more closely with human rankings under MIX than under DOC.

Domain	SEG	DOC	MIX
Literary	67s	53s	41s
News	92s	78s	76s
Social	74s	57s	60s
Overall	76s	61s	56s

Table 4: Per-segment annotation time across conditions, by domain. The decrease from SEG to DOC and MIX suggests that context shapes annotators' processing, regardless of whether that context is coherent or not.

5.6 Domain-wise replication

We replicate the analyses across our three domains, ordered by an intuitive (and unmeasured) notion of context demand: literary highest, where narrative voice and character continuity carry across a document; news lowest, with standardized tone and self-contained documents; social in between. Under H1 to H3, larger DOC-vs-MIX differences should emerge where cross-sentential dependence is stronger.

Figure 4 shows that the equivalence pattern holds across all three domains, with negligible effect sizes ($|d| < 0.13$ across all paired comparisons). We observe no graded effect. H1 differences range from $+0.79$ in news (the only domain showing the predicted direction) to -1.00 in literary, with no

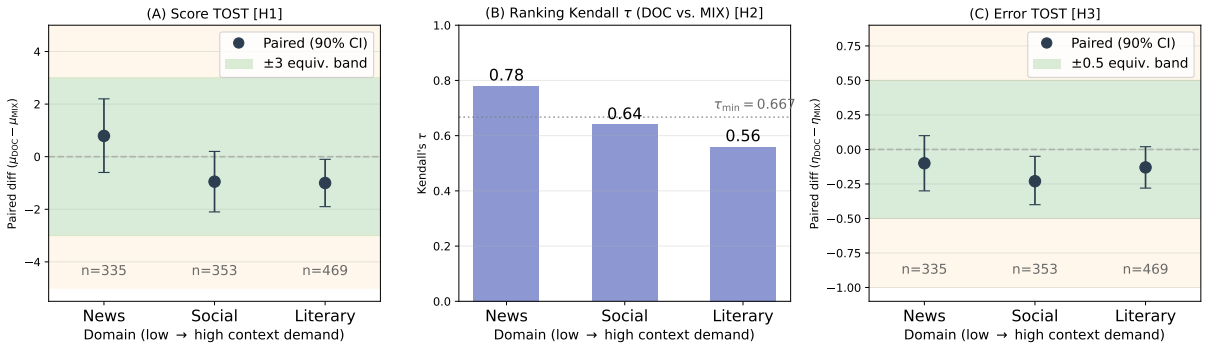

Figure 4: Domain-stratified replication of H1 to H3 on human ESA, with domains ordered by increasing context demand (news < social < literary). Throughout, the plotted paired difference is **DOC** − **MIX**. **(A)** Score TOST: paired mean differences on segment scores. All point estimates fall within the ± 3 equivalence band. **(B)** System ranking agreement between **DOC** and **MIX**. The dotted line marks the conventional acceptability threshold ($\tau_{\min} = 0.667$). **(C)** Error TOST: paired mean differences on per-annotation error counts. All estimates fall within the ± 0.5 equivalence band.

monotonic relationship (Panel A). H3 effects are largest in social rather than literary (Panel C). The one signal that aligns with the intuitive ordering is system ranking agreement (Panel B), where $\tau_{\text{DOC,MIX}}$ decreases monotonically from news (0.78) through social (0.64) to literary (0.56, below the threshold $\tau_{\min} = 0.667$). Bootstrap intervals on these values qualify the ordering. Resampling segments with replacement ($B = 10{,}000$) gives $[0.556, 0.867]$ for news, $[0.378, 0.778]$ for social, and $[0.422, 0.689]$ for literary. The three intervals overlap substantially, and a value near 0.56 falls inside all of them, and social likewise falls below $\tau_{\min}$ in the majority of resamples. The monotonic ordering in Panel B is therefore not resolvable at this sample size.[2]

We read this cautiously: literary rests on two source documents (Table 5), and a coherence-sensitivity account predicts several signals moving together, whereas here only τ moves, with the score difference within margin and reversed in direction. A single moving signal on two documents is better explained by sampling instability. What the replication establishes is that the equivalence holds in all three domains, not that context demand has no effect anywhere.

6 Conclusion

Since the ALPAC report (Hutchins, 2003), MT evaluation has been built, largely tacitly, around segment-level human judgments. Decades of gold data have accumulated under this paradigm, shaping the metrics subsequently trained or calibrated against it. Our counterfactual experiment shows what this legacy actually captures. Across human ESA and 14 automatic metrics, including discourse-targeted variants and LLM-based scorers, we find no evaluator that distinguishes coherent documents from deliberately incoherent ones, in any of the three domains tested.

Two explanations can be set aside. The first is perception: readers asked about consistency separate the two conditions reliably, and document presentation measurably changes how annotators work. The second is context length: the document-level metrics we tested received the full document and still failed, and the LLM-based scorers, with effectively unbounded context, fared no differently. What is missing is a place to put the judgment, both a notion of what constitutes document-level quality in the gold data that drives the field and an interface that elicits it.

The implication is direct. Training or fine-tuning translation models or metrics on segment-based gold is unlikely to produce document-level competence. It can, at best, reproduce the segment-level ceiling that this gold encodes. Genuine progress in document-level MT will likely require new evaluation protocols designed to elicit document-level judgments rather than merely to display document-level context, and new gold data collected under them. Until then, what is reported as document-level progress is, under our measure, hard to separate from segment-level progress.

[2] Intervals here are computed over segments rather than over source documents as in Appendix B.2, since the literary domain rests on two source documents.

Limitations

Language selection. Our investigation prioritizes depth over breadth, with 18,420 annotations on a single language pair where document-level signals are strongest. Korean's discourse-sensitive features (pro-drop, honorifics, cross-segment entity tracking) make it a setting where ESA, if it registers coherence at all, should do so most readily. A null in this most-favorable setting bounds what can reasonably be expected elsewhere, but it does not establish the same result elsewhere. Our finding is scoped to English→Korean under ESA, and replication across typologically diverse language pairs is the primary next step rather than a formality.

Domain composition. The three domains are unevenly represented: literary contributes 44.6% of all annotations but rests on only 2 source documents (Table 5). Domain-stratified results (Section 5.6) should therefore be read as replication checks rather than as a domain comparison, and the literary ranking result in particular is reported with bootstrap intervals for this reason.

Sample sizes and IAA. Our analyses are based on 10 annotators and 9 MT systems, yielding $\binom{9}{2} = 36$ pairwise comparisons for ranking statistics. Krippendorff's α values are modest ($\alpha < 0.25$ across all conditions), reflecting both the subjectivity of fine-grained translation quality assessment and the small annotator pool. We mitigate small-sample limitations with within-annotator paired designs, bootstrap confidence intervals, and statistical clustering (Appendix B.2).

Scope of the counterfactual. Our MIX condition breaks document coherence by combining segments from multiple MT systems, a strong but single form of incoherence. What it perturbs is consistency: register, terminology, and entity form. It leaves source content and sentence order intact, and it cannot produce errors internal to a single system's output, such as a mistranslated anaphor or a wrong pro-drop resolution. MIX is thus best understood as an upper bound on the consistency signal rather than as a sample of naturally occurring document-level errors. The manipulation checks (Table 6) rule out a weak manipulation as an explanation of the equivalence, though other explanations remain open. Whether ESA would register subtler perturbations, such as random reordering or gradual degradation, is untested.

Scope of contribution. We focus on ESA, a current standard protocol in WMT-style MT evaluation. Whether the same failure extends to other protocols, such as MQM with explicit discourse-level error categories, is an open question. We diagnose the failure but do not propose a replacement; designing an interface that elicits document-level judgments is itself a research program.

Acknowledgment

This work was supported by the G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317); the Ministry of Science and ICT (MSIT), and the National IT Industry Promotion Agency (NIPA) through the Advanced GPU Utilization Support Program (02-26-01-0499). This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT) (IITP-2026-RS-2022-00156360).

Ethics Statement

Our study involves expert human annotators. Participation was voluntary, and annotators were compensated at a fair market rate of $40 per hour. All annotations were anonymized, and no personally identifiable information was collected.

Licenses of artifacts. All datasets, models, and software used in this study are publicly available for research purposes. The WMT25 test set is distributed under CC-BY 4.0; the automatic metrics xCOMET-XXL, METRICX-24-XXL, chrF, and doc-COMET are released under Apache 2.0. The LLM-as-judge metrics GEMBA V2 and FALCON are computed via the OpenAI API (GPT-4.1) under OpenAI's terms of service. Pearmut is available under MIT. Our 18,420 ESA annotations are released under CC-BY 4.0 at https://huggingface.co/datasets/trotacodigos/esa-counterfactual-enko, and the analysis code under MIT at https://github.com/trotacodigos/esa-counterfactual.git.

Intended use. Our use of the WMT25 test set, automatic metrics, and the ESA protocol is consistent with their intended use for MT evaluation research. The annotations and datasets we release

are intended for research use only, consistent with the access conditions of the source data.

Use of AI assistants. We acknowledge the use of AI assistants (Claude Opus 4.6) for writing refinement and code review during paper preparation. All experimental design, analysis, and claims are the authors' own.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

Sung-Kwon Choi, Gyu-Hyeun Choi, and Youngkil Kim. 2018. Automatic evaluation of English-to-Korean and Korean-to-English neural machine translation systems by linguistic test points. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Michael P. Fay and Michael A. Proschan. 2010. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. 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.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

C. A. E. Goodhart. 1984. *Monetary Theory and Practice: The U.K. Experience*. Macmillan, London.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 283–289, Paris, France.

John Hutchins. 2003. Alpac: The (in)famous report. In Sergei Nirenburg, Harold L. Somers, and Yorick A. Wilks, editors, *Readings in Machine Translation*. The MIT Press.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2025. GEMBA v2: Ten judgments are better than one. In *Proceedings of the Tenth Conference on Machine Translation*, pages 926–933, Suzhou, China. Association for Computational Linguistics.

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.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Ahrii Kim. 2025a. Context is ubiquitous, but rarely changes judgments: Revisiting document-level MT evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 81–97, Suzhou, China. Association for Computational Linguistics.

Ahrii Kim. 2025b. Falcon: Holistic framework for document-level machine translation evaluation.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. 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. 2024a. 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, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

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.

Minjae Lee, Youngbin Noh, and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646, Abu Dhabi, UAE. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Arien Mack and Irvin Rock. 1998. *Inattentional Blindness*. The MIT Press.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Wafaa Mohammed and Vlad Niculae. 2025. Context-aware or context-insensitive? assessing LLMs' performance in document-level translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 126–137, Geneva, Switzerland. European Association for Machine Translation.

Wafaa Mohammed, Vlad Niculae, and Chrysoula Zerva. 2026. Unlocking latent discourse translation in

LLMs through quality-aware decoding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4752–4774, Rabat, Morocco. Association for Computational Linguistics.

Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Dojun Park and Sebastian Padó. 2024. Multi-dimensional machine translation evaluation: Model evaluation and resource for Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11723–11744, Torino, Italia. ELRA and ICCL.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):4381.

Dmitry Popov, Vladislav Negodin, Ekaterina Enikeeva, Iana Matrosova, Nikolay Karpachev, and Max Ryabinin. 2025. Refined assessment for translation evaluation: Rethinking machine translation evaluation in the era of human-level systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22079–22095, Suzhou, China. Association for Computational Linguistics.

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

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. Evaluating metrics for document-context evaluation in machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 812–814, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Donald J. Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.

Daniel J. Simons and Christopher F. Chabris. 1999. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074.

Yixiao Song, Parker Riley, Daniel Deutsch, and Markus Freitag. 2025. Enhancing human evaluation in machine translation with comparative judgement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20536–20551, Vienna, Austria. Association for Computational Linguistics.

Yirong Sun, Dawei Zhu, Yanjun Chen, Erjia Xiao, Xinghao Chen, and Xiaoyu Shen. 2025. Fine-grained and multi-dimensional metrics for document-level machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 1–17, Albuquerque, USA. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal. European Association for Machine Translation.

Anne M. Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kuang-Da Wang, Shuoyang Ding, Chao-Han Huck Yang, Ping-Chun Hsieh, Wen-Chih Peng, Vitaly Lavrukhin, and Boris Ginsburg. 2025. Extending automatic machine translation evaluation to book-length documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32323–32339, Suzhou, China. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Billy Wong and Chunyu Kit. 2010. The parameter-optimized ATEC metric for MT evaluation. In *Proceedings of the Joint Fifth Workshop on Statistical*

Machine Translation and MetricsMATR, pages 360–364, Uppsala, Sweden. Association for Computational Linguistics.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models? *Transactions of the Association for Computational Linguistics*, 13:1789–1811.

Vilém Zouhar and Tom Kocmi. 2026. Pearmut: Human evaluation of translation made trivial. *Preprint*, arXiv:2601.02933.

Appendix

A Experimental Setup

A.1 Dataset

Source data. The Conference on Machine Translation (WMT) has organized shared tasks for MT research since 2006 (Koehn and Monz, 2006). The General Machine Translation shared task, WMT's central event, releases parallel test sets across multiple language pairs and domains each year, against which both academic and industrial systems are evaluated under standardized human and automatic protocols. Test documents are curated from authentic sources (news articles, social media, literature, speeches) and translated independently by all participating systems, with professional reference translations provided. The multi-system parallel format enables direct comparison of system outputs on identical source material, which our counterfactual evaluation design requires. We use the WMT 2025 edition (Kocmi et al., 2025), which is the most recent and explicitly targets document-level evaluation.

Statistics. Table 5 summarizes the final dataset statistics by domain. The 10 systems comprise 9 MT systems (CLAUDE-4, GPT-4.1, GEMINI-2.5-PRO, GEMTRANS, YOLU, WENYIIL, LANIQO, IRB-MT, COMMANDA-MT) and one human reference (👤REF). The domain composition is unbalanced, reflecting WMT 2025's natural distribution. We address this through per-domain replication and length normalization, confirming that aggregate findings hold within each domain and across document lengths.

For statistical comparisons across conditions (Section 5), we restrict analysis to segments evaluated under all three conditions, yielding $n = 5{,}920$ annotations per condition. For paired analyses, the number of paired observations varies by comparison: $n = 308$ (**DOC** vs. **MIX**), $n = 364$ (**DOC** vs. **SEG**), and $n = 1{,}249$ (**SEG** vs. **MIX**). The variation reflects different overlap structures of (annotator, segment) tuples across condition pairs.

Preprocessing. We apply the following pipeline to prepare segments for human evaluation. We first prioritize segments expected to yield the most informative human judgments based on their automatic evaluation scores, following SUBSET2EVAL (Zouhar et al., 2025), and select the documents that contain them as whole documents. Segments shorter than

	News	Social	Literary	Total
Source documents and segments				
# documents	13	8	2	23
# unique segments	80	90	137	307
Mean segs/doc	6.2	11.3	68.5	13.4
System outputs				
# MT systems	9	9	9	9
# human references	1	1	1	1
# unique translations	800	900	1,370	**3,070**
Annotations				
# evaluations	2,400	2,700	4,110	9,210
# total annotations	4,800	5,400	8,220	**18,420**

Table 5: Dataset statistics by domain. *Mean segs/doc* is # unique segments divided by # documents, counted after the sub-document splitting described in Appendix A.1. Each unique translation is evaluated under three conditions (SEG, DOC, MIX) and annotated by two independent annotators, yielding 18,420 total annotations. Note the unbalanced composition: the literary domain rests on 2 source documents but contributes 44.6% of all annotations.

10 tokens are appended to their immediately preceding segment, with paragraph boundaries (\n\n) preserved throughout. The 23 documents vary substantially in length, particularly between literary documents (long-form narrative) and news or social documents (short articles or posts). To reduce length-driven variance in document-level evaluation, we split longer documents into sub-documents that approximate the cross-domain mean length. Segments exceeding the domain-average token length (80 tokens in the WMT setting) are split at the nearest line break (\n) to preserve source–target alignment. If no line break exists within the segment, the segment is retained at its original length. Each document in the resulting corpus contains at least two segments, so that document-level context is available for evaluation.

A.2 Construction of MIX

Sampling procedure. For a source document d with segments $s_1, s_2, \ldots, s_n$, the DOC condition presents all n segments translated by a single system. The MIX condition uses the same n source segments in the same order, but each segment translation is randomly drawn from the pool of system outputs for that segment. The sampling is constrained so that every system in the pool contributes at least one segment within a MIX document, and each segment prefers a system other than the one used at the same position under DOC. The resulting document preserves source content and sentence order while introducing system-level variation across consecutive segments, breaking discourse coherence by construction. We use a fixed random seed for reproducibility. Tasks are batched with a maximum of 300 segments each to balance annotator workload.

Study 1: do the pooled systems differ? The first study compares *system pairs*. To verify that the systems pooled in MIX produce distinguishable translations, we conducted a pairwise preference study on the three system pairs with the highest overlap in MQM ranks (computed by AutoRank; Kocmi et al., 2025): 👤REF vs. GEMINI-2.5-PRO (both rank 1–3), GPT-4.1 vs. CLAUDE-4 (rank 4–6 vs. 4–7), and GEMTRANS vs. WENYIIL (rank 5–10 vs. 5–12). If raters can reliably discriminate quality for these maximally overlapping pairs, systems with larger rank separation should be at least as distinguishable.

For each pair, we sampled one document per domain (54 source segments total). Each item presented the source alongside two anonymized translations in randomized order. Three undergraduate linguistics majors independently labeled each item as "A is better," "B is better," or "No difference," producing 486 judgments. We report both raw results and domain-normalized results (sample-size-weighted averaging with a 50% cap), since the literary domain accounts for 78% of segments.

We assess discriminability primarily through pooled tie rates (the proportion of "No difference" judgments), with per-rater tie rates as a transparency check on consistency. We additionally apply a two-sided sign test to non-tie judgments for directional preference, though this is secondary to the discriminability question.

Table 6 reports the results. Across all three pairs, raters perceived a quality difference in 85–94% of judgments, with per-rater tie rates converging across the three raters (GPT-4.1 vs. CLAUDE-4: 9.3/0.0/9.3%; GEMTRANS vs. WENYIIL: 22.2/3.7/11.1%; 👤REF vs. GEMINI-2.5-PRO: 16.7/11.1/16.7%) and all three ranking the pairs in the same order of discriminability. The sign test is borderline for 👤REF vs. GEMINI-2.5-PRO ($p = .050$) and non-significant elsewhere, but it targets directional preference rather than discriminability: a non-significant result implies that raters did not consistently favor one system, not that they could not distinguish the two.

Since even the maximally overlapping pairs are distinguishable in at least 85% of judgments, the systems pooled in MIX span a quality range broad

System Pair (AutoRank)	Split	n	m1	m2	Tie	Sign Test	Distribution
REF vs. GEMINI-2.5-PRO (1–3 / 1–3)	Overall (Raw)	162	35.2	50.0	14.8	$p = .050$	
	Overall (Normalized)	162	38.5	50.0	11.5		
	Literary	126	32.5	50.0	17.5	—	
	News	15	53.3	40.0	6.7	—	
	Social	21	38.1	57.1	4.8	—	
GPT-4.1 vs. CLAUDE-4 (4–6 / 4–7)	Overall (Raw)	162	54.3	39.5	6.2	$p = .062$	
	Overall (Normalized)	162	62.7	31.3	6.0		
	Literary	126	47.6	46.0	6.3	—	
	News	15	80.0	20.0	0.0	—	
	Social	21	76.2	14.3	9.5	—	
GEMTRANS vs. WENYIIL (5–10 / 5–12)	Overall (Raw)	162	39.5	48.1	12.3	$p = .275$	
	Overall (Normalized)	162	40.3	43.8	15.9		
	Literary	126	38.9	51.6	9.5	—	
	News	15	40.0	13.3	46.7	—	
	Social	21	42.9	52.4	4.8	—	

Table 6: Pairwise preference results across three system pairs (486 judgments total; 162 per pair, 3 annotators). For each pair, we report Overall results (Raw and Normalized via domain-weighted averaging with a 50% cap) and per-domain breakdowns. Distribution bars show m1 preference, m2 preference, and tie. Sign tests apply to non-tie judgments at the overall (raw) level only.

enough that MIX combines genuinely different translations rather than near-identical ones. We stress what this does not establish. Difference is a property of the pooled systems, not of the assembled document, and a reader who can tell two translations apart need not experience their concatenation as inconsistent. The second study addresses that question directly.

Study 2: does the assembled document read as inconsistent? The second study compares *documents*. We sampled 102 passage pairs, each showing the same six contiguous source segments twice, once as DOC and once as MIX. The DOC system for a pair is always one of the systems contributing to that pair's MIX version, so the two members are approximately matched on mean segment quality and differ mainly in whether a single system produced the whole passage. Presentation order was randomized and raters were told nothing about how the passages were constructed. Three raters (researchers in the field) chose for each pair which passage reads as the work of a single consistent translator.

DOC was chosen in 89 of 102 pairs (87.3%, 95% CI $[79.4, 92.4]$), with Krippendorff's $\alpha = 0.73$ across the three raters and a two-sided binomial test against chance at $p < 0.001$. Because the two members of a pair are approximately quality-matched by construction, this preference is unlikely to be explained by MIX being worse on average. It tracks within-passage consistency.

What MIX does and does not perturb. Table 7 illustrates the mechanism. Korean marks speech level on every sentence-final verb, so a change of contributing system mid-passage is audible to a native reader. In the passages we inspected, MIX also produced terminology drift and entity-form inconsistency within a passage. Errors internal to a single system's output, such as a mistranslated anaphor or a wrong pro-drop resolution, are not created by the manipulation, but neither are they removed by it: they occur in DOC and MIX alike, as does the source content and its sentence order. The one dimension on which the two conditions differ is within-document consistency, and MIX perturbs it densely and deliberately. That is what makes the null informative: a protocol blind to this is unlikely to register the sparser forms that arise naturally.

A.3 Human Annotation

Recruitment & compensation. We worked with a recruitment vendor to identify and contract a pool of raters. After a qualification step, the vendor set up an anonymized Slack channel for ongoing communication. Prior to the main annotation, we held an onboarding session via Zoom to distribute a Korean annotation guideline and address questions. Raters were then given one week to complete a

	System	Translation	Level
***DOC**: single system throughout*			
s_5	GEMINI	...거의 같은 **높이다.**	Plain
s_6	GEMINI	...손을 흔들어 **주었다.**	Plain
s_7	GEMINI	...막지 **않는다.**	Plain
***MIX**: three systems*			
s_5	GEMINI	...거의 같은 **높이다.**	Plain
s_6	GPT	...계속 손을 **흔들었습니다.**	**Formal.p**
s_7	IRB-MT	...가리지 **않네요.**	**Informal.p**

Table 7: Three consecutive segments of one social-media document. Korean marks speech level on every sentence-final verb (bolded). **DOC** holds a single level throughout, whereas **MIX** moves Plain (한다체) → Formal-polite (합쇼체) → Informal-polite (해요체) within three sentences, a register discontinuity with no counterpart in the source. Both passages were shown in the same document-level ESA interface, and **MIX** was *not* scored lower.

(a) Segment-level

(b) Document-level

Figure 5: The segment- and document-level evaluation screens of Pearmut. For **SEG** items, raters are presented with a single segment, whereas **DOC** and **MIX** items display a full document.

warm-up annotation, after which we provided individual feedback on the balance and consistency of their annotations relative to other raters. Raters were compensated at an industry-competitive rate of $40 per hour.

Data collection. Main annotation was conducted from March to April 2026. To ensure sufficient within-rater coverage while limiting fatigue, raters were asked to annotate regularly over the one-month period, averaging approximately 1 to 2 hours per day. Progress was monitored daily for attention checks, with intervention when over- or under-pacing was detected. After collection, we screened annotations for outliers and requested re-annotation for flagged items where appropriate.

Annotation protocol. We follow the ESA protocol (Kocmi et al., 2024b), adapted for Korean. Annotators evaluate each translation along two axes: fluency and adequacy at the segment level, and discourse phenomena such as consistency, logical flow, formality, and coherence at the document level. The procedure has three steps: (i) marking error spans, (ii) labeling each span as *Major* or *Minor*, and (iii) assigning an overall score from 0 to 100, anchored at four reference points (Nonsense, Broken, Middling, Perfect) as defined by Pearmut. Figure 5 displays the segment- and document-level evaluation screens.

The annotation guideline was prepared in Korean and distributed to all annotators prior to the task, specifying the evaluation procedure, score anchors, error categories, and tool usage. Critically, annotators were not informed of the counterfactual design: **MIX** documents were presented under the same interface as **DOC**, with no indication that some documents combined outputs from multiple systems. This blinding is essential to the validity of our comparison, as awareness of the manipulation would induce task demand effects that obscure ESA's natural behavior on incoherent input.

Instructions on discourse phenomena. Blinding concerned the manipulation, not the phenomena. The guideline directed raters to discourse-level properties explicitly: its document-level section names consistency of terminology and entity forms, logical flow between sentences, formality and speech-level consistency, and overall coherence, with a worked Korean example of each. Raters were instructed to mark a span and lower the segment score when such a property is violated, exactly as they would for a fluency or adequacy error. The negative result is therefore not attributable to annotators having been left unaware that discourse phenomena were in scope.

We deliberately did not add a separate coherence-scoring item. Our claim concerns ESA *as deployed*, whose premise is that showing the full document is sufficient for discourse errors to surface in the per-segment outputs. **MIX** is dense with such errors and, under that premise, should score lower. A ded-

icated coherence question would test a different instrument, and would in effect be one of the redesigned protocols we call for in Section 6 rather than a fix to the one under test.

A.4 Automatic Metrics

We use a suite of MT evaluation metrics covering surface, learned, and LLM-as-judge approaches at both segment and document granularities.

A.5 Segment-level metrics

chrF (Popović, 2015). A character-level surface metric that computes the F-score of character n-grams between hypothesis and reference. chrF was among the strongest-performing metrics at WMT25 (Lavie et al., 2025). We use CHRF++ (adding word n-grams up to order 2) for Korean.

XCOMET-XXL (Guerreiro et al., 2024). A learned reference-based metric built on XLM-R XXL (11B parameters). It produces regression-style quality scores from (source, hypothesis, reference) triples and has been a top performer at recent WMT shared tasks. COMETKIWI-XXL (Rei et al., 2023) is its reference-free counterpart.

METRICX-24-XXL (Juraska et al., 2024). A learned metric based on mT5-XXL that predicts MQM-style quality scores. Unlike most metrics, METRICX produces error scores (0 = perfect, 25 = worst), which we negate in our analyses for consistency with other metrics. METRICX-24-QE-XXL is its reference-free variant.

GEMBA V2 (Junczys-Dowmunt, 2025). An LLM-as-judge metric. We use GPT-4.1 as the underlying LLM, matching the configuration under which GEMBA V2's performance was originally reported (Junczys-Dowmunt, 2025). The model is prompted with source-hypothesis pairs and asked to elicit a 0–100 quality rating following the GEMBA V2 prompt template, which extends the original GEMBA (Kocmi and Federmann, 2023) with error-span elicitation prior to scoring.

A.6 Document-level metrics

d-BLEU (Liu et al., 2020). Corpus-level BLEU computed on full document hypotheses concatenated against full document references. d-BLEU is sensitive to large-scale n-gram overlap rather than to discourse coherence.

doc-COMET (Vernikos et al., 2022). A document-aware extension of COMET that incorporates surrounding sentences into the (source, hypothesis, reference) triple, allowing the model to use intra-document context.

SLIDE (Raunak et al., 2023). A reference-free document-level metric (Sliding Document Evaluator) that scores translations through a sliding window over consecutive segments, aggregating per-window quality estimates from a base QE model (COMETKIWI-XXL in our experiments).

FALCON (Kim, 2025b). An LLM-as-judge metric designed for document-level MT evaluation. It prompts the LLM with full documents under nine discourse-related criteria and elicits both MQM-style error annotations and overall quality scores. We use GPT-4.1 as the underlying LLM, matching the configuration under which FALCON's performance was originally reported.

A.7 Implementation notes

We exclude the 👤REF system, since hypothesis-reference identity yields trivially perfect scores for reference-based metrics. Because segment-level metrics produce identical document-level scores for SEG and DOC by construction, we focus the metric analysis on the DOC vs. MIX comparison. Reference-based metrics (chrF, XCOMET-XXL, METRICX-24-XXL, d-BLEU, doc-COMET) use the publicly available references included in the WMT dataset, while reference-free metrics (COMETKIWI-XXL, METRICX-24-QE-XXL, SLIDE, GEMBA V2, FALCON) operate on (source, hypothesis) only. For all metrics except METRICX-24-XXL variants, higher scores indicate better quality; METRICX-24-XXL produces error scores (lower = better), which we negate in correlation and ranking analyses to maintain consistent direction across metrics.

Equivalence test (H1) details. We set the equivalence margin per metric scale: 3 points for 0–100 scales (chrF, d-BLEU), 0.05 for 0–1 scales (XCOMET-XXL variants), and 0.5 for the 25-point METRICX-24-XXL scale. All metrics support equivalence at TOST $p < 0.001$.

Restriction to segment-level metrics for human↔metric agreement. The human-metric agreement analysis is restricted to segment-level metrics. Document-level metrics produce one score per document, which under DOC can be attributed to the single system that translated the

document but under MIX corresponds to documents combining segments from multiple systems. Distributing a single MIX document score across constituent systems requires modeling assumptions (such as proportional to segment count) that introduce analysis artifacts, so we restrict the agreement analysis to segment-level metrics, where each segment carries an unambiguous system label.

A note on token truncation. We apply XCOMET-XXL and METRICX-24-XXL variants at the document level for completeness. However, both models have maximum input lengths shorter than typical document lengths in our data. XCOMET-XXL accepts up to 512 tokens, while METRICX-24-XXL accepts up to 1,536 tokens. Documents are therefore truncated under both metrics, but the effect is most severe for XCOMET-XXL-doc which yields very low scores (~ 0.16 in both conditions), compared with the typical XCOMET-XXL range of 0.5–0.9 on segment-level inputs. METRICX-24-XXL-doc variants are less affected, owing to the longer input window. We include all document-level results for completeness but interpret XCOMET-XXL-doc as reflecting truncation rather than genuine document-level evaluation.

B Statistical Methodology

B.1 Equivalence Testing (TOST)

The Two One-Sided Tests (TOST) procedure formally tests the null hypothesis $H_0 : |\mu_1 - \mu_2| \geq \epsilon$ against the alternative $H_1 : |\mu_1 - \mu_2| < \epsilon$, where ϵ is the smallest difference of practical interest (Fay and Proschan, 2010). A significant TOST result supports the conclusion that the two conditions are practically equivalent within $\pm\epsilon$. Unlike a non-significant conventional test, which cannot establish equivalence, TOST shifts the burden of proof. Equivalence must be demonstrated.

Equivalence margin. We set $\epsilon = 3$ points on the 100-point ESA scale on two grounds. First, it corresponds to approximately 0.2 standard deviations of the empirical score distribution ($\sigma \approx 13$–15). Second, it lies below the typical magnitude of IAA in ESA evaluation. Equivalence at $\epsilon = 3$ implies equivalence at any larger margin. We additionally report results at $\epsilon \in \{5, 7\}$ in Table 8 for transparency.

Paired and independent TOST. The main results report *paired TOST*, which matches observations on the same segment evaluated under both conditions, controlling for between-segment variance. We additionally report *independent-sample TOST* as a robustness check (Table 8). It treats observations as independent samples and yields consistent conclusions across all comparisons.

B.2 Ranking Analysis [H2]

System rankings are induced from segment-level mean scores aggregated by system, over the $n = 9$ MT systems, excluding the human reference REF. This yields $\binom{9}{2} = 36$ pairwise comparisons. (Table 2 additionally displays REF for human ESA, where it is scored like any other system.) We compute Kendall's τ between ranking pairs (r_a, r_b) for conditions $a, b \in \{\text{SEG}, \text{DOC}, \text{MIX}\}$, preferring τ over Spearman's ρ given the small number of systems and of tied ranks. The threshold $\tau_{\min} = 0.667$ corresponds to moderate ranking agreement, and values below it indicate re-ordering across conditions.

Bootstrap confidence intervals. We report 95% bootstrap confidence intervals computed by resampling source documents with replacement ($B = 10{,}000$ resamples). Resampling is at the document level because annotations are nested within the 23

source documents and the segment-level resampling would understate the variance. For each resample we recompute system-level mean scores, induce rankings, and compute τ, taking the 2.5th and 97.5th percentiles of the resulting distribution.

For human ESA, τ(DOC, MIX) has a bootstrap mean of 0.737 (SD $= 0.097$) and a 95% CI of $[0.539, 0.911]$, and 22.5% of resamples fall below $\tau_{\min}$. Expressed in discordant pairs, the point estimate of 0.778 corresponds to 5 of 45 pairs ordered differently under the two conditions. The pairs that reverse are not arbitrary: they are the pairs whose mean scores are closest under DOC, where the difference between systems is smaller than the sampling variability of the means themselves (GPT-4.1 and WENYIIL differ by 0.05 ESA points, GEMINI-2.5-PRO and 👤REF by 0.19). Exact ties among such pairs arise in 1.1% of resamples and are handled by the τ_b correction. τ assigns a strict order to these pairs regardless, so its value is largely determined by which documents enter a given resample. We therefore treat the clustering analysis in Table 2 as the primary evidence for H2.

Statistical clustering. Following the WMT methodology (Kocmi et al., 2025), we group systems into statistically equivalent clusters via pairwise bootstrap tests ($B = 1{,}000$ resamples, $\alpha = 0.05$; the smaller B follows WMT and suffices for a significance decision, whereas the interval estimates above require more resamples in the tails). Two systems are placed in the same cluster when their mean score difference is not significantly different from zero, so that pairs of the kind described above are left unordered rather than assigned an arbitrary rank. This yields condition-specific cluster assignments that are stable under small score perturbations.

B.3 Error Annotation Analysis [H3]

We compute error counts per 1,000 target characters to control for varying segment length. Paired TOST compares η_{MIX} against η_{DOC} on a per-segment basis with margin $\epsilon_{\text{err}} = 0.5$ errors per annotation (H3). We additionally report the proportion of annotations marked as error-free (score $= 100$) and the distribution of error severity (Major/Minor) across conditions.

Positional analysis of error spans. For signal ③ we locate each span by the character offset of its midpoint within its segment, normalized to $[0, 1]$ by segment length. A span counts as *boundary-localized* if that position falls in $[0, 0.2)$ or $(0.8, 1]$, that is, within the first or last fifth of the segment. Under a null of no positional structure the expected rate is 40%, and the observed rates near 50% reflect the tendency of annotated spans to include sentence-initial and sentence-final material in all conditions alike, which is why the between-condition contrast, not the absolute level, carries the inference. Confidence intervals are bootstrap intervals over segments ($B = 10{,}000$).

Three robustness variants leave the result unchanged. (i) Excluding zero-width omission markers, which carry no span extent and are placed by convention at a segment edge. (ii) Restricting to document-interior segments, since the first and last segment of a document have only one neighboring seam. (iii) Restricting to *Major* spans, on the reasoning that a register or terminology break should be marked as major if it is marked at all. In every variant the MIX $-$ DOC difference remains within ± 1.5 points with a confidence interval covering zero.

B.4 Behavioral Measure

IAA. We compute Krippendorff's α on the interval scale at two units of analysis. At the document level, the unit is $\langle$doc_id, system$\rangle$, with per-annotator scores averaged across the constituent segments. At the segment level, the unit is $\langle$doc_id, seg_id, system$\rangle$.

Annotation time. Per-segment annotation time is computed as the cumulative time between consecutive annotation actions within a task, excluding gaps longer than 120 seconds (treated as inactive periods), divided by the number of segments in the task. This definition isolates active evaluation effort from incidental pauses.

Comparison	Design	n	Mean diff.	90% CI	TOST p-value		
					ε=3	ε=5	ε=7
DOC vs. MIX	independent	5920/5920	-0.82	[-1.25, -0.40]	<0.001	<0.001	<0.001
	paired	308	-1.00	[-2.05, +0.04]	<0.001	<0.001	<0.001
DOC vs. SEG	independent	5920/5920	+0.55	[+0.11, +0.99]	<0.001	<0.001	<0.001
	paired	364	+0.07	[-0.83, +0.97]	<0.001	<0.001	<0.001
SEG vs. MIX	independent	5920/5920	-1.37	[-1.79, -0.95]	<0.001	<0.001	<0.001
	paired	1249	-1.09	[-1.65, -0.52]	<0.001	<0.001	<0.001

Table 8: TOST equivalence test results across all pairwise condition comparisons, designs (independent vs. paired), and equivalence margins $\varepsilon \in \{3, 5, 7\}$ points on the 100-point ESA scale. For independent designs, n is reported as n_1/n_2; for paired designs, n is the number of matched pairs. All eighteen cells support equivalence at $p < 0.001$, demonstrating robustness to comparison choice, study design, and equivalence margin.

C Detailed Results

C.1 Score Equivalence [H1]

TOST across pairwise comparison and metrics. Table 8 reports TOST results for all three pairwise comparisons (DOC vs. MIX, DOC vs. SEG, SEG vs. MIX) under both independent and paired designs. Independent and paired analyses yield the same qualitative conclusion for every comparison, with paired estimates consistently closer to zero. For DOC vs. SEG, the mean difference shifts from $+0.55$ under the independent design to $+0.07$ under the paired design. For SEG vs. MIX, it shifts from -1.37 to -1.09. Part of the apparent between-condition variance is therefore attributable to inter-annotator differences rather than to genuine condition effects.

Robustness to margin choice. We report results at equivalence margins $\epsilon \in \{3, 5, 7\}$ in Table 8. All eighteen cells (three comparisons $\times$ two designs $\times$ three margins) yield $p < 0.001$, with mean differences ranging from -1.37 to $+0.55$. The equivalence conclusion holds across comparison choice, study design, and margin value within the range tested.

C.2 IAA

Document- vs. segment-level agreement. Krippendorff's α depends on the unit of analysis. Table 9 reports both units (see Appendix B.4 for definitions). At the document level, DOC leads with $\alpha = 0.186$. At the segment level, the order reverses, with MIX leading ($\alpha = 0.225$) and DOC lowest ($\alpha = 0.155$). The pattern suggests that under DOC, annotators incorporate document context, producing diverse per-segment judgments while converging at the document level on shared overall impressions. Under MIX, the absence of coherent document context returns judgments to segment-only assessment, yielding tighter segment-level agreement.

Condition	Krippendorff's α	
	Per document	Per segment
SEG	0.149	0.172
DOC	**0.186**	0.155
MIX	0.128	**0.225**

Table 9: Inter-annotator agreement, computed at two units. *Per document* treats ⟨doc_id, system⟩ as the item, averaging each annotator's segment scores within it; *per segment* treats ⟨doc_id, seg_id, system⟩ as the item. The orderings invert between the two: DOC agrees most per document but least per segment. All values fall below the conventional threshold for acceptable reliability ($\alpha \geq 0.667$), consistent with known variability in scalar MT evaluation (Graham et al., 2013; Kocmi et al., 2025).

A note on MIX's document-level agreement. For MIX, the document-level item ⟨doc_id, system⟩ does not correspond to a single physical document. Each MIX document is a unique recombination of segments from multiple systems, and the same ⟨doc_id, system⟩ pair may refer to different segment sets across annotators. The lower document-level α for MIX is therefore largely a structural artifact rather than a behavioral signal. Our analysis emphasizes the segment-level α comparison, where item identity is unambiguous.

Absolute reliability remains low. All values fall below the conventional threshold for acceptable reliability ($\alpha \geq 0.667$), in line with known variability in scalar MT evaluation. This does not compromise our primary findings. The analyses operate on aggregated scores derived from large per-condition samples ($n = 5{,}920$ each), where annotator-level noise is averaged out, and within-annotator paired analyses in Section 5.1 explicitly control for inter-

Metric	Condition	Kendall's τ	p	Pearson's r
Surface				
chrF	DOC	0.667	0.013	0.815
	MIX	0.889	< 0.001	0.946
d-BLEU	DOC	0.667	0.013	0.831
	MIX	0.889	< 0.001	0.950
Learned ref-based				
XCOMET-XXL	DOC	0.500	0.075	0.642
	MIX	0.611	0.025	0.803
METRICX-24-XXL	DOC	0.556	0.045	0.604
	MIX	0.667	0.013	0.844
Reference-free (QE)				
COMETKIWI-XXL	DOC	0.222	0.477	0.252
	MIX	0.333	0.260	0.505
METRICX-24-QE-XXL	DOC	0.278	0.359	0.278
	MIX	0.389	0.180	0.588
LLM-as-judge				
GEMBA V2	DOC	0.504	0.075	0.652
	MIX	0.722	0.007	0.872

Table 10: System-level human-metric correlations by condition ($n = 9$ systems). Kendall's τ values for QE metrics (COMETKIWI-XXL, METRICX-24-QE-XXL) are not statistically significant at $p < 0.05$, reflecting the small system pool and narrow score distributions of these metrics. We focus on the relative pattern across conditions in our analysis.

	Scale	Range	SD	Min	Max
Human evaluation	0–100	12.28	4.09	78.60	90.88
Surface metrics					
chrF	0–100	9.81	3.08	23.37	33.18
d-BLEU	0–100	11.66	3.64	15.52	27.18
Learned reference-based					
XCOMET-XXL	0–1	0.076	0.032	0.716	0.792
METRICX-24-XXL *	0–25	1.030	0.382	3.826	4.856
Reference-free (QE)					
COMETKIWI-XXL	0–1	0.056	0.016	0.806	0.861
METRICX-24-QE-XXL *	0–25	0.839	0.268	3.602	4.441
LLM-as-judge					
GEMBA V2	0–100	8.43	2.71	79.52	87.95

Table 11: System-level mean score distributions across nine systems under each metric. Range is computed as max − min of system means. Reference-free QE metrics produce the narrowest output ranges ($\sim$ 3–6% of metric scale), making ranking-based human agreement noise-sensitive. Surface metrics span ranges comparable to human evaluation. *METRICX variants are error scores (lower = better); ranges are reported in the metric's native scale.

annotator variance.

Per-annotator normalization. Annotator-wise z-score normalization shifts the absolute values but preserves the ordering across conditions. After normalization, document-level α for DOC rises from 0.186 to 0.23, while SEG and MIX remain around 0.15 and 0.13 respectively. The document-level result is therefore not an artifact of annotator-specific scoring styles.

C.3 Human–Metric Agreement

We provide detailed per-metric results. Kendall's τ and Pearson's r at $n = 9$ are therefore noise-sensitive, and we emphasize relative patterns across conditions rather than absolute correlation values.

Per-metric, per-condition correlations. Table 10 reports detailed human-metric system correlations for both DOC and MIX conditions, complementing Section 5.4. For surface metrics (chrF, d-BLEU), agreement shifts from $\tau = 0.67$ under DOC to $\tau = 0.89$ under MIX. The same direction holds for learned metrics at lower absolute values. XCOMET-XXL shifts from 0.50 to 0.61, and METRICX-24-XXL from 0.56 to 0.67. Reference-free QE variants show the same upward shift from a lower baseline, from $\tau = 0.22$–0.28 under DOC to 0.33–0.39 under MIX. The LLM-as-judge metric GEMBA V2 follows the same pattern, moving from $\tau = 0.50$ under DOC to 0.72 under MIX —matching surface metrics in the magnitude of its MIX shift while sitting between learned reference-based and surface metrics in absolute agreement.

On the narrow output range of QE metrics. Table 11 reports system-level score distributions. Surface metrics (chrF, d-BLEU) span $\sim 10\%$ of their scale, comparable to human evaluation ($\sim 12\%$), and GEMBA V2 spans a similar $\sim 8\%$. Learned reference-based metrics are narrower ($\sim$ 4–8%), and reference-free QE metrics are narrowest of all: COMETKIWI-XXL at $\sim 5.6\%$ (range $= 0.056$) and METRICX-24-QE-XXL at $\sim 3.4\%$ (range $=$ 0.84 on a 25-point scale). With distributions this narrow, ranking-based agreement becomes noise-sensitive, and small score fluctuations can flip ranks—plausibly contributing to QE metrics' lower absolute correlation with humans.