

PragmaSpan at WMT26: Taxonomy-Guided Few-Shot Error Span Detection and an Empirical Analysis of MPP

Soyoung Lee

BrainVentures

sylee@brainventur.com

Abstract

We present PragmaSpan, a training-free, few-shot GPT-5.5 error-span detector with a seven-axis diagnostic taxonomy for WMT26 Task 1, wrapped in a reliability-oriented harness using incomplete-response-aware retries, deterministic parsing, offset recovery, exact-span deduplication, and independent artifact validation. The taxonomy provides a domain-sensitive diagnostic representation: its interpersonal and discourse axes track domain conversationality and match gold errors at least as often as its traditional axes, with no significant detection-F difference from a standard Multi-dimensional Quality Metrics (MQM) category tree. We use the system to characterize the severity-weighted Match with Partial overlap and Partial credit (MPP) metric. Across prompt and model comparisons, human-agreement calibration, and post-hoc refinement, we find (i) prompt variants do not differ significantly on our sample, while GPT-4.1-to-GPT-5.5 gains exceed differences among variants; (ii) human-human MPP is low on our multi-annotator surrogate sample, and system-human agreement is not significantly different from this empirical baseline; (iii) improvements in matched-pair severity or boundary accuracy can coincide with lower MPP, demonstrating the need to interpret MPP jointly with matched-pair diagnostics and match coverage; and (iv) a calibration gain on a single-annotator MQM benchmark does not replicate on a multi-annotator Error Span Annotation benchmark. Interpreting error-span improvements therefore requires complementary metrics and annotation-regime checks.

1 Introduction

WMT26 continues the unified Automated Translation Quality Evaluation track that WMT25 established by consolidating the prior separate QE and metrics shared tasks (Lavie et al., 2025). Task 1 asks systems to localize translation errors as

character-level spans with a minor/major severity label, scored with Match with Partial overlap and Partial credit (MPP), extended with severity weighting (minor=0.25, major=1.0), one-to-one optimal matching, and micro-averaging, via the reference implementation `span-mt-metrics-eval` (Perrella et al., 2026). We use our shared-task system to examine human agreement and the behavior of practical interventions under this scoring configuration. Because the task includes dialogue and document-context domains, we also examine whether discourse- and interpersonal-oriented axes add diagnostic value beyond conventional accuracy and fluency categories.

Our system is architecturally simple: no fine-tuning, no retrieval, and one successful inference per segment (with retries only for API failures or incomplete output), deliberately so that our ablations isolate properties of the *metric* and *task* rather than interactions among elaborate modeling components. We report negative and non-generalizing results alongside gains, using them to distinguish what aggregate MPP, matched-pair diagnostics, and match coverage reveal about system changes and why gains on one benchmark may not transfer to another.

Our contributions are fourfold: (i) a reference-optional, taxonomy-guided detector that extends traditional accuracy- and fluency-oriented error categories with axes for discourse, speaker stance, social relations, and document-context consistency, yielding a domain-sensitive diagnostic representation with no statistically significant difference in detection F from the standard MQM category tree on our two benchmarks; (ii) a reliability-oriented harness that preserves the boundary between model diagnosis and artifact construction through fixed retries, span normalization, and independent artifact validation; (iii) an empirical analysis of severity-weighted MPP that relates score differences to human agreement and shows why aggregate MPP,

matched-pair diagnostics, and match coverage provide complementary evidence about system improvements; and (iv) a cross-benchmark account of successful, negative, and non-generalizing interventions, including a statistically grounded decision to rely on the organizers’ automatic Task 1→2/3 conversion.¹

2 System Description

2.1 Taxonomy and prompt

Our approach builds on prompted fine-grained MT evaluation, exemplified by AutoMQM (Fernandes et al., 2023) and GEMBA-MQM (Kocmi and Federmann, 2023), and most directly follows the latter’s fixed, language-agnostic template. Trained approaches include xCOMET’s span tagging (Guerreiro et al., 2024) and GemSpanEval’s fine-tuned generative error annotation (Juraska et al., 2025). We instead use a fixed few-shot prompt without task-specific training. We depart from the GEMBA-MQM template in three ways. First, and most centrally, we replace the standard MQM category tree with a custom seven-axis taxonomy, motivated by WMT26 Task 1’s inclusion of dialogue and document-context domains in which interpersonal and discourse-structural phenomena are not well served by MQM’s primarily accuracy/fluency-centric categories; this substitution is a diagnostic and interpretive layer on top of the same underlying span-detection behavior, not a claim of a different detection mechanism, and §5.4 directly measures what it does and does not buy. Second, like GEMBA-MQM, we use one fixed prompt across language pairs, but replace its three canonical single-error demonstrations with four multilingual examples (detailed below). Third, unlike GEMBA-MQM’s fixed reference-free (quality-estimation) design, we make reference use conditional rather than fixed (also detailed below).

The model is asked to diagnose each candidate translation along seven axes: *semantic accuracy*, *terminology consistency*, and *target-language naturalness* are genre-agnostic and apply to any translation. The remaining four axes—*discourse function*, *speaker stance*, *register/social relation*, and *document-context consistency*—are motivated by dialogue- and document-context translation. They capture interpersonal or discourse-structural phenomena but are not limited to conversational text:

rhetorical contrast, epistemic stance, and coreference also occur in monologic genres. §5.4 reports how often the model predicts these axes across domains. Each detected issue carries one axis, one severity (minor/major), and zero or more character-offset spans into the candidate translation (an empty span list marks an omission, or a cross-sentence consistency issue within a multi-sentence segment, such as our speech-domain items; neither case has one exact textual locus). The taxonomy is not part of the official scoring (the metric consumes only span offsets and severity, and the category field is explicitly ignored by the reference scorer), so its role is confined to structuring the model’s diagnosis and feeding the local alternative Task 2/3 conversions evaluated in §2.4.

The final prompt (v2) combines coverage-oriented prompting with a reference-optional, fixed multilingual design. Its fourth few-shot example demonstrates that a single segment can and should yield multiple independent issues (dense-error calibration), while an explicit clause-by-clause instruction discourages the model from stopping after the first detected error (Table 3, v1→v2). The prompt incorporates a reference translation when one is available while retaining reference-free operation otherwise. Its multilingual design grounds the taxonomy’s dialogue-pragmatic dimensions in concrete Korean phenomena, including honorifics, formality levels, and addressee choice, while retaining one shared template: we use the same prompt for English-to-Indonesian and all other evaluated language directions, without language-pair-specific prompt tuning.

The complete prompt template and four fixed demonstrations are reproduced in Figures 1 and 2 in Appendices A and B.

2.2 Model

We use GPT-5.5 with `reasoning_effort=none`. Final inference starts with an output-token limit of 8,192; retry adjustments are described in §2.3. Development first selected GPT-5.5 over GPT-4.1 based on our two surrogate benchmarks (§3); a final current-code comparison against GPT-5.6 Sol found opposite directions on ESA and MQM and therefore retained the cross-benchmark-stable GPT-5.5 (§4.1). We did not fine-tune or ensemble; §5 reports why several natural extensions did not survive testing.

¹Code: <https://github.com/8pril/pragmaspan>

2.3 Reliability-oriented inference and submission harness

API-based structured evaluation can fail at two levels: semantic diagnosis and artifact construction. Even an otherwise usable response may fail to yield a valid submission record because of truncation, malformed JSON, misaligned offsets, or duplicate spans. We isolate the latter with a three-stage harness: inference detects and retries incomplete responses; normalization reconciles and deduplicates spans; and submission assurance enforces exact coverage and schema validity. Because these checks concern output structure and span consistency, retry and subsequent processing follow fixed rules; model calls remain stochastic, but a given response produces the same auditable outcome without an additional model-based revision of the diagnosis.

Transport-level incomplete responses trigger up to three total attempts, with the output-token limit doubled on each retry. A completed response with a malformed JSON tail instead undergoes conservative repair: the parser retains only fully closed issue elements and records the intervention.

Model issues are normalized against the candidate string before serialization. A span is retained when its offsets and reported text agree; otherwise, exact text is relocated to the nearest occurrence when unambiguous. If the text cannot be found but integer offsets are available, the offsets are clamped and the candidate substring is used; empty or ambiguous spans are rejected. Exact duplicate ranges are collapsed at serialization, retaining major severity on disagreement, while semantic_accuracy issues left with no valid span after normalization populate the official omission field. Appendix C gives the complete algorithm and worked examples.

Submission assurance uses two independent code paths. The assembler checks coverage while constructing the nested submission; a separate validator, without invoking the assembler, reloads the artifact and official input to verify exact item/system coverage and span, severity, omission, and duplicate constraints. Any discrepancy prevents finalization. These safeguards affect decoding and submission integrity, not the model’s diagnostic decision rule; we therefore do not attribute benchmark score gains to the harness.

In the final blind run, 8,416 of 10,327 rows triggered at least one intervention, driven primarily

Mechanism	Rows	Events
Incomplete-response retry	115	134
Conservative JSON repair	116	135
Offset recovery	8,408	32,369
Ambiguous-span rejection	0	0
Invalid-span rejection	7	11
Exact-span deduplication	576	718
Severity-conflict promotion	131	141

Table 1: Blind-run harness interventions. Counts measure operational interventions, not estimated MPP gains.

by reconciling model-provided offsets with exact candidate substrings. The separately executed final validation confirmed complete coverage of 397 items and all 10,327 item/system predictions in the finalized artifact.

2.4 Task 2/3 strategy

We submit Task 1 only, relying on the organizers’ announced automatic conversion into Tasks 2/3 rather than opting out in favor of our own local conversion strategy; §6 reports the replication and comparison that justifies this choice.

3 Experimental Setup

Since the official test set is blind, development-time evaluation uses two surrogate benchmarks scored with the official metric and reference implementation:

- **ESA** (Error Span Annotation) (Kocmi et al., 2024) (WMT25 General MT human evaluation, surrogate gold, closely resembling the actual test conditions). Each segment–system pair has two annotations; we score both their union and each annotation separately because the union can double-count overlapping marks, while either individual annotation omits errors found only by the other annotator. The dataset covers literary, speech, social, news, and educational domains across En–X/X–Y directions. Since WMT aggregation conventions have changed across years (Lavie et al., 2025) and the WMT26 consolidation policy was unavailable, we report both views.
- **MQM** (Multidimensional Quality Metrics) (Freitag et al., 2021) (WMT22 General MT, en–de, Google’s public MQM release). Its 21,040 segments have one professional annotation each (verified programmatically), providing a single-convention benchmark at the

Model	MQM F	ESA F (merged)	ESA F (per-ann.)
GPT-4.1	0.279	0.280	0.232
GPT-5.5	0.341	0.328	0.239

Table 2: Model comparison, fixed prompt (v2).

cost of coverage limited to en-de and four domains (conversation, e-commerce, news, social). We re-check every positive MQM finding on ESA before treating it as general (§5.3).

Model experiments use fixed samples of 100 ESA and 200 MQM segment-system pairs, drawn with sampling seed 42. ESA is stratified jointly by language pair, error-free label, and human-score bin (< 40 , $40 \leq 70$, $70 \leq 85$, ≥ 85); MQM is stratified by error-free label. Sampling shuffles records within each stratum and draws in round-robin order across strata until the target size is reached. The full-corpus human-human agreement analysis is reported separately in §4.3.

All comparisons use the official MPP implementation with one-to-one optimal matching, micro-averaging, and the announced severity weights.

4 Results

4.1 Model Comparison

GPT-5.5 gains $+0.062$ F on MQM, $+0.048$ on ESA merged gold, and $+0.007$ per annotator over GPT-4.1 (Table 2). The first two gains exceed the spread among the prompt variants explored during development with GPT-4.1 (§4.2); the per-annotator gain is smaller but directionally consistent.

Before the blind run, we repeated GPT-5.5 versus GPT-5.6 Sol model selection using the final inference and exact-span deduplication code. All three paired-bootstrap comparisons (Sol minus GPT-5.5, 5,000 resamples) include zero: ESA merged $0.320 \rightarrow 0.333$ ($+0.013$, 95% confidence interval (CI) $[-0.017, +0.047]$), per annotator $0.229 \rightarrow 0.243$ ($+0.015$, CI $[-0.006, +0.036]$), MQM $0.338 \rightarrow 0.319$ (-0.019 , CI $[-0.074, +0.034]$). Because the apparent ESA gain did not transfer across benchmarks, we retain GPT-5.5 rather than claim a reliable Sol improvement.

4.2 Prompt ablation (fixed model: GPT-4.1)

Qualitatively, v3’s recall gain was outweighed by its precision loss and it began flagging stylistic non-issues in human reference translations. The v4 guardrail did not reduce reference flagging (5/5 segments, as for every other variant). We did not fully adopt v5 because gold span granularity varied from single characters to full paragraphs within the same annotator pool, undercutting its premise.

Paired bootstrap 95% confidence intervals (5,000 resamples) for the differences in merged-gold F (Table 3) were: v2 without reference minus v1, $+0.016$ $[-0.009, +0.043]$; v2 with reference minus v2 without reference, $+0.013$ $[-0.016, +0.043]$; v3 minus v2 without reference, -0.004 $[-0.033, +0.022]$; and v4 minus v2 without reference, -0.003 $[-0.027, +0.024]$. All intervals include zero. We selected v2 based on its favorable point estimate and the absence of the qualitative regressions observed in later variants, not on statistically established superiority.

4.3 Calibrating against human agreement

We additionally score human annotators against each other with the same official metric (one ESA annotator’s list as “prediction,” the other’s as “reference”). Human-human agreement is low: $F=0.315$ over the full 54,419-pair WMT25 corpus and $F=0.208$ on our 100-sample subset, indicating substantial variability in span selection and boundaries. The system’s $F=0.239$ is statistically indistinguishable from this same-sample baseline (difference $+0.032$, paired-bootstrap 95% CI $[-0.034, +0.096]$; probability of no improvement 0.169) (Koehn, 2004). This empirical baseline is not a formal ceiling, but its low level cautions against interpreting small system differences as reliable quality differences.

5 Practical Limits of the Evaluated Interventions

5.1 Oracle decomposition of remaining error

Table 4 shows that even after jointly correcting boundaries and severities on matched pairs (a local one-to-one, intersection-over-union (IoU)-maximizing alignment, a proxy for the official MPP matching), absolute gaps of 0.490 on MQM and 0.506 on ESA remain relative to a perfect score: the oracle only corrects already-matched pairs, so it neither adds spans for missed errors nor removes

Variant	Prompt configuration	ESA F (merged)
v1	Base prompt, 3 few-shot examples	0.250
v2 (submitted)	+ dense-error example and “be exhaustive” instruction	0.267 (no ref) / 0.280 (with ref)
v3	Removed the leniency clause to increase recall	0.263
v4	v2 + anti-overflagging guardrail and clean Example E	0.264
v5	Root-cause span consolidation	—

Table 3: Prompt ablation with GPT-4.1.

	MQM	ESA
Base (GPT-5.5)	0.341	0.328
+ oracle span boundaries	+0.068	+0.075
+ oracle severity (both dir.)	+0.087	+0.072
+ both	+0.169	+0.166
Residual F gap (detection)	0.490	0.506

Table 4: Oracle decomposition with detection fixed. Each “+” row reports the absolute F increase over the base score after correcting boundaries, severities, or both on locally matched pairs. Residual F gap is $1 - (F_{\text{base}} + \Delta F_{\text{both}})$: the absolute distance from a perfect score that remains attributable to *which* spans are predicted at all.

Variant	F	Sev. acc.	Bound. acc.	Matches
Base	0.341	0.706	0.376	85
Self-refine	0.335	0.706	0.388	85
Cross-model refine	0.326	0.753	0.420	81

Table 5: Post-hoc refinement results on MQM. F is the official MPP score; severity accuracy, exact-boundary accuracy, and the match count are computed on the official one-to-one MPP matching itself.

unmatched false positives. The residual is therefore a detection gap relative to this consolidated gold, whose interpretation is complicated by the annotation variability quantified in §4.3.

5.2 Why matched-pair diagnostics can disagree with MPP

We tested whether a second, “refine” pass over first-pass detections could recover some of the boundary/severity headroom above, with two variants: **Self-refine** (GPT-5.5 reviews its own first-pass issues, free to edit spans and severity) and **Cross-model refine** (Gemini-2.5-Pro edits GPT-5.5’s first-pass spans/severity). Table 5 reports both.

Both refiners improve at least one matched-pair accuracy measure, but neither improves total MPP. Self-refine keeps the same 85 matches as the base system and leaves severity accuracy unchanged. Although exact-boundary accuracy rises, official precision/recall fall from 0.269/0.465

to 0.266/0.450 because partial-overlap credit decreases on other pairs across the sample and the predicted-severity mass in MPP’s precision denominator increases (Appendix D works through one such pair). Exact-boundary accuracy counts exact matches but does not capture the magnitude of overlap changes across pairs; more exact matches can therefore coexist with lower aggregate overlap credit. Cross-model refine also loses four official matches (85→81), with precision/recall falling further to 0.266/0.421. Matched-pair accuracy is conditional on the pairs an optimal matching happens to keep; it does not characterize the total severity-weighted overlap that MPP actually scores, so the two can move in opposite directions. Post-hoc span refinement should therefore be evaluated with official MPP, matched-pair diagnostics, and match coverage together: MPP captures aggregate effects, the diagnostics identify which local properties changed, and the count exposes lost matches.

5.3 Span-locked severity calibration: a benchmark-generalization failure

Motivated by §5.2, we designed a calibration step that fixes span locations and the issue set by construction: a second pass is shown each first-pass issue’s span (read-only) and asked to re-judge *only* its severity; spans, axes, and the issue set are copied through unconditionally regardless of what the model returns. We further note that because MPP severity credit is $\min(\text{pred_weight}, \text{gold_weight})$, under-graded corrections (gold=major, predicted=minor) raise a matched pair’s credit directly, giving a realistic oracle ceiling from detection quality alone of +0.054 (MQM) / +0.051 (ESA). Corrections of over-graded errors (gold=minor, predicted=major) leave the pair’s matching credit unchanged. However, lowering the predicted severity reduces MPP’s segment-level precision denominator and therefore raises precision. This denominator effect explains why the aggregate oracle gains of +0.087/+0.072 reported in §5.1 exceed the credit-gain ceilings

Variant	MQM F	Sev. acc.
Base: single pass	0.341	0.702
A: Gemini, severity only	0.329	0.702
B: GPT-5.5, higher-reasoning severity only	0.359	0.737
C: GPT-5.5, higher-reasoning joint pass	0.304	0.711

Table 6: Three-way test on MQM. “Sev. acc.” is matched-pair severity accuracy under local $\text{IoU} \geq 0.5$ diagnostic matching. C’s detection density also fell from 1.02 to 0.83 spans/segment.

above. Corrections of over-graded errors still matter for the Task 2/3 penalty conversion (§2.4), where predicted severity directly sets the penalty.

Variant B (Table 6) recovers roughly a third of the realistic severity headroom (+0.019 of +0.054); this gain has a paired-bootstrap 95% CI of $[-0.010, +0.047]$ and a 0.098 probability of no improvement. C has slightly higher matched-pair severity accuracy than the baseline, but lower prediction density and MPP, again illustrating that better matched-pair diagnostics need not translate into a higher official score.

However, re-running variant B on ESA (a different annotation convention and domain mix) found no improvement: merged-gold F decreased slightly (0.328→0.326) and diagnostic matched-pair severity accuracy worsened (0.683→0.647); a small per-annotator F uptick (+0.011, 0.239→0.250) coincided with no underlying severity-accuracy gain and appears attributable to incidental re-matching rather than genuine calibration. Because B’s MQM gain was not statistically significant and did not replicate consistently on ESA, we omit the severity-only second pass, following our cross-benchmark selection criterion (§3).

5.4 What the taxonomy actually contributes

To isolate taxonomy effects, we ran GPT-5.5 on both benchmarks using the standard MQM top-level categories and equivalently relabeled demonstrations, while holding the rest of v2 fixed. Differences were small and non-significant: MQM 0.341 vs. 0.334 (diff +0.006, 95% CI $[-0.044, +0.058]$) and ESA merged 0.328 vs. 0.316 (diff +0.012, 95% CI $[-0.027, +0.054]$), in both cases favoring our taxonomy. This directly supports our characterization of the taxonomy in §2.1 as diagnostic organization rather than a distinct detection mechanism.

A finer-grained ablation that removes the four interpersonal and discourse axes before scoring fur-

ther localizes the null result: F changes by at most ± 0.001 on either benchmark, and these axes account for only 2.5–6.4% of predicted spans, tracking domain conversationality (speech 10.8% > literary 7.6% > social 6.3% > news 3.4% > edu 3.0% on ESA). On ESA, they also match gold errors at least as often as the three traditional axes (28.2% vs. 23.8% at $\text{IoU} \geq 0.5$); the same comparison reverses on MQM (20.0% vs. 28.1%), but only five predicted spans used a pragmatic axis there, too few to interpret directionally. We therefore treat the taxonomy as a domain-sensitive diagnostic representation that supports a density-robust aggregation layer, with little effect on aggregate detection F in the evaluated domain mix (§6 reports the downstream conversion analysis).

6 Task 2/3 Conversion and Submission Scope

The organizers’ automatic Task 1→2/3 conversion applies MQM-style penalties (major: -10 , minor: -2.5). On our GPT-5.5 outputs for the 100-segment ESA sample, it achieved Pearson $r = 0.497$, Spearman $\rho = 0.425$, and Task 3 MCC = 0.262. Our *diagnosis_only* alternative derives each segment score from one severity judgment per taxonomy axis rather than from the number or length of emitted spans. Among our local conversion strategies, only *diagnosis_only* maintained stable Task 2 correlation as prediction density increased from GPT-4.1 to GPT-5.5; span-count-based conversions degraded sharply. This best local alternative did not differ significantly from the organizers’ conversion in a paired bootstrap with 5,000 resamples (organizer – ours: r : $\Delta = +0.076$, 95% CI $[-0.046, +0.216]$; ρ : $\Delta = +0.016$, CI $[-0.082, +0.115]$; MCC: $\Delta = -0.003$, CI $[-0.203, +0.190]$). For MCC, each side’s tuned threshold was fixed across resamples, so the interval excludes threshold-selection uncertainty. We therefore submitted Task 1 only and relied on automatic conversion.

Our runs cover the official high-priority English→Indonesian direction (331 items; 10,261 item/system predictions) and the English→Korean QEBreak challenge set (66 predictions). English→Korean was removed from the official roster because human evaluation could not be secured. Official per-language-pair leaderboard results were not available at the system-paper deadline.

7 Discussion: Interpreting Error-Span Improvements

Our experiments expose two distinct obstacles to interpreting error-span improvements: aggregate and component-level measures can move in opposite directions, and the evaluation target can vary across annotators. Aggregate MPP obscures how graded overlap credit, severity weighting, match coverage, and prediction mass contribute to score changes. Matched-pair severity and exact-boundary accuracy help attribute local changes, but condition on retained matches and omit aggregate costs elsewhere. Improved matched-pair diagnostics do not establish overall improvement when match coverage or prediction mass worsens; conversely, lower MPP does not imply that every detector component deteriorated. Error-span detection therefore cannot be adequately characterized by a single aggregate score. To attribute score changes, we therefore report matched-pair severity and exact-boundary accuracy alongside retained-match counts to track match coverage, and inspect MPP precision/recall and prediction mass when the aggregate score and diagnostics diverge.

MPP determines the official ranking, whereas interpreting what changed requires matched-pair severity and boundary diagnostics, total overlap credit, match coverage, and prediction mass. This decomposition, however, does not resolve uncertainty in the annotation target itself. Low human-human MPP reveals substantial variation in error selection and boundary placement, so a small system-score difference may reflect closer agreement with a particular annotation realization rather than a reliable improvement in error detection. The failure of the MQM calibration gain to replicate on ESA further shows that conclusions may depend on the annotation regime. We therefore treat small or regime-specific gains cautiously and do not regard them alone as reliable evidence of improved error-span detection.

8 Conclusion

We submit a training-free, few-shot LLM span detector wrapped in a reliability-oriented harness, addressing artifact failures that benchmark scores alone do not expose. Its taxonomy provides a domain-sensitive diagnostic representation. Our taxonomy-based local Task 2 conversion remained stable as prediction density increased but did not differ significantly from the organizers' conversion,

so we submitted Task 1 only. The evaluated prompt, calibration, and taxonomy interventions did not establish consistent MPP gains across our experiments. Instead, our analysis shows that a single aggregate score cannot explain whether localization, severity, match retention, or prediction mass improved. System comparisons should therefore report official MPP alongside matched-pair diagnostics and match coverage, and interpret the results against human agreement and across annotation regimes. PragmaSpan thus provides both an operational span-evaluation system and evidence that error-span evaluation requires aggregate scoring and decomposed diagnostics.

Limitations

- **Multimodal blind spot.** 88 English→Indonesian items have video+ASR (speech) and 82 have screenshots (social); we include the ASR transcript as context but cannot inspect video or screenshots, even though the ESA protocol for spoken content rates against the video, not the transcript alone.
- **No explicit neighboring-segment context.** Although the blind set contained other segments from the same document, we did not provide their source text or candidate translations to the model; speech items received only the accompanying ASR transcript as external context, and news, social, and software items received none. Multi-sentence items still allow intra-item consistency checks, but on single-sentence, non-speech items, `document_context_consistency` firings may instead reflect ordinary source-candidate consistency rather than genuine document-level phenomena.
- **Unvalidated artifact-construction heuristics.** Our surrogate gold does not label omissions, so the derived omission field remains unevaluated. We also did not isolate the frequency or precision effect of offset-only recovery when reported span text could not be found.
- **Single closed-source primary model.** Our final system uses GPT-5.5 with no open-weight replication, limiting independent reproducibility of the exact numbers.

- **Benchmark-to-test-distribution gap.** Both ESA and MQM are surrogates for WMT26’s actual contrastive-ESA test protocol (Pearmut tool), which may differ in annotation density and conventions; §5.3 demonstrates sensitivity to such annotation-regime differences.

Reproducibility

Appendices A and B provide the complete submitted prompt template and all four fixed demonstrations. The main text reports the model and decoding configuration, benchmark construction, official scoring configuration, post-processing rules, and statistical procedures used for the experiments.

Acknowledgments

This work was supported by the Ministry of Culture, Sports and Tourism of Korea and the Korea Creative Content Agency (KOCCA) under the 2026 Leading Company-Linked Mutual Growth Support Program (Project No. E2026-0083100001), in collaboration with NAVER Cloud; by the Ministry of Science and ICT of Korea and the National IT Industry Promotion Agency (NIPA) under the 2026 SaaS Development and Commercialization Support Program (Project No. PJT-26-040471); and by Goyang City and the Goyang Industry Promotion Agency (GIPA) under the 2026 Goyang-type Private Investment-Linked Technology Startup Support Program (Goyang TIPS).

References

- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. [The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. 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.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. [MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, Suzhou, China. 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. 2024. [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. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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 Sindhuja, 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.
- Stefano Perrella, Eric Morales Agostinho, and Hugo Zaragoza. 2026. [Span-level machine translation meta-evaluation](#). *Preprint*, arXiv:2603.19921.

A Complete Task 1 Prompt Template

The full system and user prompt template is shown in Figure 1; per-row values replace the braced slots.

```
(System) You are an expert machine translation evaluator.
Return only valid JSON matching the requested schema. Be strict, calibrated,
and avoid over-penalizing harmless variation. Evaluate translations as situated
utterances in document context, not as isolated word substitutions.

(User) Analyze the candidate translation and detect translation errors using the same
contextual-pragmatic taxonomy as the scoring system.
Represent errors as issues. Each issue belongs to one taxonomy axis, has one
severity, and may have zero, one, or multiple target-language spans as evidence.

Use character offsets into the candidate translation. The end offset is exclusive.
If an issue is an omission or document-level problem with no exact target span,
return an empty spans list for that issue. If there are no issues, return an
empty issues list.

Source language: {SOURCE_LANGUAGE}
Target language: {TARGET_LANGUAGE}
Language pair: {LANGUAGE_PAIR}

Use this contextual-pragmatic taxonomy:
- semantic_accuracy: preserved propositional meaning, including omissions and hallucinations.
- discourse_function: preserved speech act or rhetorical role, such as request, warning, refusal, joke, contrast, concession, or
emphasis.
- speaker_stance: preserved confidence, hesitation, politeness attitude, emotional force, and interpersonal distance.
- register_social_relation: appropriate formality, honorifics, address terms, pronouns, and social relation cues in the target
language.
- terminology_consistency: correct and consistent names, domain terms, glossary items, units, and named entities.
- document_context_consistency: consistency with prior context, coreference, tense/aspect, topic flow, and document-level facts.
- target_language_naturalness: grammaticality, idiomaticity, punctuation, and native-like flow.

For each axis, assign severity "none", "minor", "major", or "critical".
Use "critical" for errors that make the segment unusable, unrelated, or dangerously misleading.
For English-to-Korean, be especially sensitive to honorifics, formality level,
subject/object omission, addressee choice, names/titles, and speech-act force.

[FOUR FIXED FEW-SHOT EXAMPLES INSERTED HERE -- see Appendix B]

Annotation style:
- Prefer the smallest target span that contains the error evidence.
- Mark minor errors too, but do not annotate harmless paraphrases.
- Be exhaustive: after finding one issue, keep checking the rest of the segment clause by clause for additional, independent issues
instead of stopping early. Segments with many small imperfections should receive many separate issues, following Example D.
- If one issue is supported by multiple target spans, put all spans under the same issue.
- If the target omits source content, create an issue with an empty spans list.
- Use "major" only when meaning, usability, instruction following, or reader understanding is substantially affected.

[IF THE ITEM CARRIES A TRANSLATION INSTRUCTION:]
Translation instruction given to the system being evaluated:
{TRANSLATION_INSTRUCTION_TEXT}
Violations of this instruction (persona, formatting rules, terminology,
register constraints, required conventions) count as translation errors
and should be annotated like any other issue, on the taxonomy axis that
fits best.

Document context:
{DOCUMENT_CONTEXT, OR "(none)" IF NOT PROVIDED}

Source:
{SEGMENT_SOURCE_TEXT}

Candidate translation:
{SEGMENT_HYPOTHESIS_TEXT}

Reference translation, if available:
{SEGMENT_REFERENCE_TEXT, OR "(not provided)"}

Return JSON:
{"issues": [{"issue_id": "E1", "axis": "one taxonomy axis", "severity": "minor|major", "explanation": "short reason", "spans":
[{"start": int, "end": int, "text": "span text"}]}]}
```

Figure 1: The complete system and user prompt template used for every submitted Task 1 row. Braced items are filled per row. The four fixed demonstrations in Figure 2 are inserted at the marked position.

B Four Examples Used for Few-Shot Prompting

Example A

Source language: English

Target language: Chinese

Source:

The package will arrive tomorrow.

Candidate translation:

包裹昨天会到达。

Output:

```
{
  "issues": [
    {
      "issue_id": "E1",
      "axis": "semantic_accuracy",
      "severity": "major",
      "explanation": "'tomorrow' is translated as 'yesterday', reversing the time meaning.",
      "spans": [
        {
          "start": 2,
          "end": 4,
          "text": "昨天"
        }
      ]
    }
  ]
}
```

Example B

Source language: English

Target language: Korean

Source:

I am ready.

Candidate translation:

준비 됐어요.

Output:

```
{
  "issues": [
    {
      "issue_id": "E1",
      "axis": "target_language_naturalness",
      "severity": "minor",
      "explanation": "Korean spelling is incorrect; the natural spelling is '됐어요'.",
      "spans": [
        {
          "start": 3,
          "end": 4,
          "text": "됐"
        }
      ]
    }
  ]
}
```

Example C

Source language: English

Target language: Chinese

Source:

I will see you at practice.

Candidate translation:

我会在练习时见到你。

Output:

```
{
  "issues": []
}
```

Example D

Source language: English

Target language: Korean

Source:

Thanks for coming to the workshop, Professor Kim. I really appreciated your

presentation, but I couldn't attend the Q&A session afterward.

Candidate translation:

워크숍 와줘서 고마워, 김. 발표 진짜 좋았어. 근데 그다음 세션은 못 갔어.

Output:

```
{
  "issues": [
    {
      "issue_id": "E1",
      "axis": "register_social_relation",
      "severity": "major",
      "explanation": "Informal banmal speech is used toward a professor/superior instead of the expected honorific register.",
      "spans": [
        {
          "start": 4,
          "end": 11,
          "text": "와줘서 고마워"
        }
      ]
    },
    {
      "issue_id": "E2",
      "axis": "register_social_relation",
      "severity": "minor",
      "explanation": "The honorific title '교수님' is dropped; addressing a professor by bare name is too informal.",
      "spans": [
        {
          "start": 13,
          "end": 14,
          "text": "김"
        }
      ]
    },
    {
      "issue_id": "E3",
      "axis": "speaker_stance",
      "severity": "minor",
      "explanation": "The casual intensifier '진짜' and clipped tone read as overly familiar rather than professionally appreciative.",
      "spans": [
        {
          "start": 16,
          "end": 25,
          "text": "발표 진짜 좋았어"
        }
      ]
    },
    {
      "issue_id": "E4",
      "axis": "terminology_consistency",
      "severity": "minor",
      "explanation": "'Q&A session' is genericized to 'the next session', losing the specific meaning that it was a question-and-answer session.",
      "spans": [
        {
          "start": 30,
          "end": 36,
          "text": "그다음 세션"
        }
      ]
    }
  ]
}
```

Note how this single segment yields four separate issues across three axes: dense error segments should get correspondingly dense issue lists, not one representative issue.

Approximate English glosses for reader accessibility (not shown to the model). A: “The package will arrive yesterday.” B: “I am ready” (with a spelling error in Korean). C: “I will see you at practice.” D: “Thanks for coming to the workshop, Kim. Your presentation was really good, but I could not attend the next session.” (informal register).

Figure 2: The four fixed examples used for all language directions.

C Span Normalization Algorithm

Algorithm 1 gives the exact per-span repair procedure summarized in §2.3; it is applied independently to every span of every model-reported issue before flattening.

Algorithm 1 Span offset normalization (one reported span)

Require: reported span (s, e, t) ; candidate string c

```

1:  $validInts \leftarrow (s \text{ and } e \text{ are both integers})$ 
2:  $original \leftarrow t \neq \varepsilon$  and  $validInts$  and  $0 \leq s \leq e \leq |c|$  and  $c[s:e] = t$ 
3: if  $t \neq \varepsilon$  and not  $original$  then
4:    $O \leftarrow$  start indices of every exact occurrence of  $t$  in  $c$ 
5:   if  $O \neq \emptyset$  then
6:     if  $s$  is a usable integer then
7:        $s \leftarrow \arg \min_{o \in O} (|o - s|, o)$  ▷ closest to reported  $s$ ; ties broken leftmost
8:     else if  $|O| = 1$  then
9:        $s \leftarrow O_1$ 
10:    else
11:      return AMBIGUOUS ▷ discard: no location hint to disambiguate
12:    end if
13:     $e \leftarrow s + |t|$ ;  $validInts \leftarrow true$  ▷ else:  $O = \emptyset$  – leave  $(s, e)$  untouched and fall through below
14:  end if
15: end if
16: if not  $validInts$  then
17:   return INVALID
18: end if
19:  $s \leftarrow \max(0, \min(|c|, s))$ ;  $e \leftarrow \max(s, \min(|c|, e))$ 
20:  $t' \leftarrow c[s:e]$ 
21: if  $t' = \varepsilon$  then
22:   return INVALID
23: end if
24: return  $(s, e, t')$ , status = (UNCHANGED if  $original$  else REPAIRED)

```

Surviving spans are then flattened across issues, sorted by $(start, end, issue_id)$, and exact-duplicate $(start, end)$ groups are collapsed to a single entry, keeping major over minor on disagreement. The official omission field is computed separately, from the pre-flattening issue list, as the maximum severity among semantic_accuracy issues left with no valid span after normalization (null if there are none).

Worked examples. (i) A model reports `text="foo"` at offset (7, 10) for candidate `"foo bar foo"` (`candidate[7:10]=" fo"`, so repair triggers); `"foo"` occurs at offsets 0 and 8. Since $|8 - 7| = 1 < |0 - 7| = 7$, the span is repaired to (8, 11), not (0, 3): the occurrence closer to the model's own offset wins. (ii) A model reports `text="xyz"` (absent from the candidate) at offset (3, 6) on the 5-character candidate `"abcde"`: no occurrence exists, so (3, 6) is clamped to (3, 5) and repaired to `text="de"` – the offsets are trusted over the unmatchable text.

D A Refinement Example Where MPP Falls

An MQM example shows how self-refine lowers MPP without losing the match or changing severity. The gold and base prediction are identical major spans at [37, 47), *an die Tür*, with $P = R = F = 1$. Self-refine widens the prediction to [11, 47), *Holen Sie sofort jemanden an die Tür*, while retaining major severity. The overlap remains 10 characters but prediction length grows to 36, yielding $P = 10/36 = 0.278$, $R = 10/10 = 1$, and $F = 2PR/(P + R) = 0.435$. The match and severity remain correct, but the wider boundary lowers pair-level F—an individual overlap loss underlying the aggregate result in §5.2, where exact-boundary accuracy increased despite a lower overall MPP score.