

FACET at WMT 2026 Automated Translation Quality Evaluation Task

Ahrii Kim¹ Chanjun Park^{2*} Seong-heum Kim^{1,3*}

¹AI-Bio Convergence Research Inst. ²School of Software ³Dept. of Intelligent Semiconductors

Soongsil University

{ahriikim,chanjun.park,seongheum}@ssu.ac.kr

Abstract

Different error types in machine translation require different evidence. Whether meaning is preserved can be judged only against the source, while whether the target is well-formed, or whether it names one entity consistently, can be judged from the target alone. We present FACET, our reference-free submission to the WMT26 Automated Translation Quality Evaluation Task, which decomposes evaluation into Fluency, Accuracy, and Consistency passes and gives each pass only the context its error type requires. A single fixed model is prompted three times, and the merged error spans yield the three task outputs, error spans, quality scores, and error-free labels, with no trained components. We also submit FACET-C, which omits the Consistency pass. Without gold labels, we characterize the predictions of FACET. Its system rankings place post-edited human translation first, and the Consistency pass changes about a tenth of segment scores while leaving the ranking nearly unchanged.

1 Introduction

Judging the quality of a translation is not one decision but several, and they do not rest on the same evidence. Whether the target preserves the meaning of the source can be decided only from the source and the target together. Whether the target is well-formed is a property of the target as a piece of language, and whether a recurring name or term keeps one rendering is a property of the target across a whole document. The source contributes nothing to the last two questions, and a single segment is too short to answer the third. LLM-based evaluators typically put all of these questions to the model in one prompt over one input (Kocmi and Federmann, 2023a; Fernandes et al., 2023; Lu et al., 2024; Junczys-Dowmunt, 2025), so every

judgment is made from the same context whether or not it is the context that judgment needs.

FACET (Fluency, Accuracy, and Consistency Evaluation of Translations) separates the judgments and the inputs. Its three passes follow the MQM error taxonomy (Lommel et al., 2013). Accuracy and Fluency are the two top-level MQM dimensions, with terminology, style, and locale conventions folded into Fluency as target-only judgments. Consistency is an MQM subtype of Fluency that we promote to a pass of its own, because it is the one Fluency judgment that cannot be made within a segment.

The passes run under asymmetric input. The Accuracy pass reads the source and the target. The Fluency and Consistency passes read the target only, each at the scope its error type requires, the segment for Fluency and the whole document for Consistency. The Consistency pass judges only recurring items, that is, named entities, domain terms, and fixed expressions. It checks that one entity does not appear under two or more names a reader would take as different referents, and that shortened or pronominal references agree with the form they refer back to. Because it judges internal consistency rather than correctness, it needs no source, and the restriction to recurring items limits overlap with the other two passes.

FACET is our submission to the WMT26 shared task on Automated Translation Quality, which merges the metrics and quality estimation (QE) tracks of previous years. The task provides reference translations as an optional input and scores segments in context, where the context is a document or a long multi-segment unit. FACET uses no references and treats the document context as an input in its own right. The system has no trained or fine-tuned components. A single fixed model is prompted, and the behavior of the system is determined by the decomposition into passes and by the input each pass receives. One merged prediction

*Corresponding authors.

 <https://github.com/trotacodigos/facet.git>

yields the three shared-task outputs, error span annotation, quality scoring, and error-free detection.

Our primary submission, FACET, uses all three passes. Whether a document-level judgment pays off under segment-level scoring is an open question, so we also submit FACET-C, an ablation that omits Consistency, and use the pair to measure what the document-level pass contributes.

2 Related Work

LLM-as-judge evaluation. Prompting large language models (LLMs) for translation quality assessment is now common practice. GEMBA (Kocmi and Federmann, 2023b) prompts GPT models for direct quality scores, and GEMBA-MQM (Kocmi and Federmann, 2023a) extends this to reference-free MQM-style error annotation (Lommel et al., 2013) with severity labels. AutoMQM (Fernandes et al., 2023) elicits error spans and categories through prompting. EAPrompt (Lu et al., 2024) combines chain-of-thought with error analysis and improves segment-level accuracy, where earlier prompting methods were weak. Rubric-MQM (Kim, 2025b) identifies label biases and the difficulty of assessing near-perfect translations, and improves reliability through rubric-style multidimensional prompting. Several later systems refine this approach in different ways. MQM-APE (Lu et al., 2025) adds an automatic post-editing stage that filters errors with no effect on quality. Rubric-as-Experts (Xu et al., 2026) allocates MQM sub-type spaces per instance and reports that broader evaluation spaces raise coverage at the cost of false positives, while narrower spaces improve precision. GEMBA V2 (Junczys-Dowmunt, 2025), built on the same GPT-4.1-MINI backbone we use, scores each segment ten times and aggregates the results to reduce variance. All of these systems judge a segment with a single prompt. FACET instead separates fluency, accuracy, and consistency into distinct passes, each given only the context its error type requires.

Quality estimation metrics. Quality estimation (QE) scores a translation from the source and the target alone, without a reference. This is the setting the present task inherits from the QE track. Learned metrics have led this line for years. COMET-QE (Chimoto and Bassett, 2022) and TransQuest (Ranasinghe et al., 2020) won the earliest WMT QE-as-a-metric subtasks, and CometKiwi (Rei et al., 2022) won the WMT22 QE task by jointly

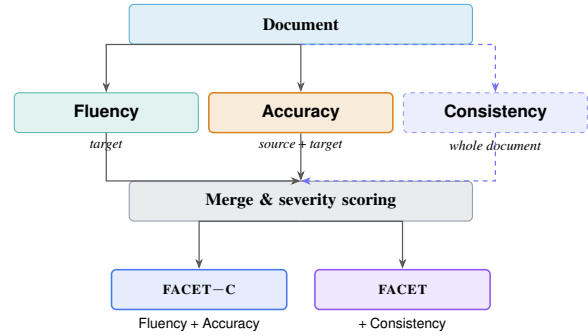


Figure 1: FACET evaluates a translation with three independent prompting passes under **asymmetric input**. Accuracy reads the source and the target, while Fluency and Consistency read the target only. Predictions are merged and scored per segment. FACET uses all three passes, and FACET-C drops Consistency (dashed).

predicting sentence- and word-level quality. More recent metrics add finer-grained outputs. xCOMET (Guerreiro et al., 2024) couples a learned metric with error-span prediction, and MetricX (Juraska et al., 2024) scores translations with or without a reference from a single model. These metrics are trained on human quality annotations such as the MLQE-PE and MQM corpora. FACET stays in the reference-free QE setting but uses no trained components, obtaining error spans, scores, and an error-free decision from prompting alone.

3 System Description

FACET evaluates a hypothesis translation with three prompting passes that run independently and are combined afterward. Each pass returns a list of error spans. A span carries a quoted region of text, the erroneous substring within it, a severity, a category, and a short reason. Figure 1 shows the pipeline. The passes differ in the input each one receives, and this difference fixes the scope at which each pass operates. Prompts are given in Appendix A.

3.1 Asymmetric Input

The three passes do not receive the same input. Accuracy is the only judgment that requires both the source and the target, since whether a translation preserves meaning cannot be decided from the target alone. Fluency and Consistency require the target only. Fluency asks whether a segment is well-formed in the target language, a property of the target as a piece of writing. Consistency asks whether a recurring entity or term keeps one rendering across the document, a property of the target across its own length. Neither question is an-

swered by the source, and supplying it would invite the model to import meaning-level concerns that belong to the **Accuracy** pass. The same asymmetry fixes the scope of each pass. **Fluency** and **Accuracy** operate on the segment, while **Consistency** operates on the whole document, the only scope at which the renderings of a term can be compared against one another.

3.2 Input Construction

The shared task provides text in paragraph-sized units. Line breaks in the data mark subtitle-cue boundaries rather than sentence boundaries, so we re-segment the source and the target into sentences with a rule-based multilingual segmenter before the passes run. Newlines are replaced with spaces in a length-preserving substitution, so the character span of each sentence maps directly back onto the original text and all offset recovery stays in the original coordinates. Presenting a segment as explicit sentences also gives the model a stable structure to iterate over, which reduces its tendency to skip material in long inputs (Kim, 2025a). For the **Consistency** pass, segments sharing a document identifier are grouped and ordered to reconstruct the full document.

3.3 Fluency Pass

The **Fluency** pass reads the target segment alone and acts as a strict proofreader of the target language. It judges spelling, grammar, agreement, word order, punctuation, orthography, register, and terminology by the norms of the target language itself. The source is withheld because a **Fluency** judgment concerns the target as language, and showing the source invites meaning-level concerns that belong to **Accuracy**.

3.4 Accuracy Pass

The **Accuracy** pass reads the source and the target together and judges whether the target preserves the meaning of the source. It covers mistranslation, wrong term, addition, hallucination, omission, and untranslated text. Source-side errors, namely omission and defective source, are recorded on the source so that content dropped from the translation is captured even when the target reads fluently. Hallucinated content unsupported by the source is assigned the highest severity.

3.5 Consistency Pass

The **Consistency** pass reads the whole target document and judges only internal consistency, never correctness, which is why it needs no source. It first identifies the recurring items whose renderings must stay stable, that is, named entities, domain terms, and fixed expressions, and records the forms in which each appears across the document. An item with a single form throughout is consistent by construction and is left to the **Accuracy** pass. An item with two or more distinct forms is checked on two points, whether the forms would read as different referents, and whether shortened or pronominal references agree in gender and number with the form they refer back to, where the target language marks such agreement.

When an item does diverge, the error is not attached to the first or the last occurrence. The pass decides from the document context alone which form does not fit the rest of the document and marks every occurrence of that form, wherever it appears. Occurrences of the form that fits are left unmarked.

3.6 Severity and Scoring

Every pass rates each error on one shared four-point scale, appended identically to all three prompts. Following the severity criteria in the human evaluation guidelines provided by WMT, scale 1 is a slight wording change with no impact on clarity or intent, scale 2 alters wording while the message stays mostly clear, scale 3 noticeably distorts meaning so that the reader is misled, and scale 4 is a substantial distortion that leaves the translation unfaithful. The prompt records the raw value, and the mapping onto Minor and Major is applied afterward. Following the correspondence in the guidelines, we cut at 3, so scales 1–2 are Minor and scales 3–4 are Major.

Before scoring, **Consistency** errors at scale 1 are discarded. On morphologically rich target languages the **Consistency** pass flags ordinary case, gender, or number inflection of a recurring term as an inconsistency, and these false positives concentrate at scale 1. Removing them leaves the higher-severity divergences that reflect cross-segment errors. **Fluency** and **Accuracy** errors at scale 1 are kept.

Because the passes are prompted and collected separately, the **Fluency**, **Accuracy**, and **Consistency** predictions for a segment arrive on separate records. We merge all predictions that share

a segment before scoring rather than letting one pass overwrite another. Error spans are located in the original text by matching the quoted region returned by the model and then the erroneous substring within it, rather than by trusting model-reported offsets. The segment score is the negated sum of the severities of its surviving errors across all passes, so a segment with errors of scale 2, 1, and 3 scores -6 and a segment with no error scores 0. A higher score is a better translation, and a segment scores 0 exactly when no error survives.

3.7 Deriving the Task Outputs

One merged, scored prediction yields all three shared-task outputs. For error span detection, each surviving error contributes its character offsets, its Minor or Major label under the cut, and its category. Consistency errors whose referents span multiple non-contiguous positions may lack a single offset, and these are omitted from the span output. For quality scoring, the segment score above is used directly. For error-free detection, a segment is labeled error-free when its score is 0, that is, when no error survives the merge.

4 Experimental Setup

Model. All three passes use a single model, GPT-4.1-MINI (gpt-4.1-mini-2025-04-14), served through the OpenAI API and accessed only through prompting with no fine-tuning. It is not an ensemble and is not based on open weights, and its parameter count is not publicly disclosed. We decode deterministically at temperature 0, so the judgments of FACET are reproducible. The model is non-reasoning, so its output is the answer itself with no intermediate reasoning trace. The maximum output length is set per request in proportion to the input. For the segment-level Fluency and Accuracy passes we allow $1.5\times$ the input length plus a floor of 1,000 tokens, capped at 8,000. For the document-level Consistency pass, whose output must enumerate every tracked item, we allow $2\times$ the input plus a floor of 2,000 tokens, capped at 16,000. Requests are issued through the batch interface, one per segment for Fluency, one per segment for Accuracy, and one per document for Consistency.

Data. We evaluate on the WMT26 test set, which spans 21 language pairs and multiple domains, from official speech-sourced material to challenge sets. Each item pairs a source with the outputs of

many translation systems, and FACET scores every system output independently, yielding several hundred thousand system-segment pairs per pass. The human annotations and the official leaderboard were not released at the time of writing, so we do not report correlation with gold labels.

Speech transcripts. A large share of the data is speech-sourced. In pilot runs the Fluency and Accuracy passes flagged filler words and hesitations, such as *um* and *uh* and their target-language equivalents, as errors, and treated their differing renderings across languages as mistakes. These are transcription artifacts, not translation errors. For speech-based items we therefore append a rule instructing the Fluency and Accuracy passes to ignore fillers, hesitations, and repeated interjections. The Consistency pass is left unchanged, since fillers are not recurring entities or terms.

Submissions. Our primary submission, FACET, uses all three passes with the scale-1 filtering of Consistency errors described in Section 3. To isolate what the document-level Consistency pass contributes under segment-level scoring, we also submit FACET-C, an ablation that omits Consistency and scores from Fluency and Accuracy alone.

5 Submission Analysis

Without gold labels, we characterize the predictions of FACET on their own terms, through the system ranking they induce, the effect of the Consistency pass, and one worked example.

System ranking. For each language pair we rank the systems by their mean FACET score and report the per-pair ranks in Table 1. Because languages differ in how many errors the passes surface, ranking within a language pair rather than pooling across pairs keeps language difficulty from being confounded with system quality. The resulting order follows the pattern one would expect from the systems involved. Strong general-purpose systems such as Wayfinder, Lumen, and GPT 5.5 rank consistently near the top, and small open models such as GLM 4-9B, GPT OSS 20B, and Qwen 3.5-9b rank consistently near the bottom.

The clearest pattern is in the human entries. *Human (postediting)* attains the best mean rank of any system at 1.3, placing first or second in every pair it enters, while *Human (from scratch)* sits mid-table at 13.3. One reading is that FACET rewards

System	cs-de	en-de	cs-vi	en-is	en-et	zh-ja	en-se	en-zh	en-th	en-cs	en-kk	en-ar	en-ru	cs-uk	en-lij	en-uk	en-llid	en-hy	en-be	en-id	en-zh-TW	avg
<i>Human (postediting)</i>	1	2	1	–	–	1	–	–	–	1	–	2	–	–	–	–	–	–	–	–	–	1.3
<i>Human (from scratch)</i>	15	–	15	8	9	–	11	–	–	–	–	–	–	–	20	–	15	–	–	–	–	13.3
Wayfinder	5	3	5	1	2	5	5	2	3	10	1	1	3	2	2	1	11	5	3	20	4	4.5
Lumen	9	7	3	2	4	8	3	4	1	5	4	9	6	1	1	4	4	2	1	11	6	4.5
GPT 5.5	2	6	2	3	1	3	7	10	2	4	6	3	8	3	11	6	3	3	2	3	9	4.6
UvA-MT	3	4	8	7	5	2	2	1	4	8	3	7	5	7	3	5	7	4	5	9	5	5.0
TRIVE	7	5	6	5	3	4	8	5	7	9	2	5	4	4	7	2	8	1	4	15	3	5.4
Cohere CAT+	4	1	4	14	12	10	1	3	9	2	16	8	2	5	5	3	2	18	13	1	2	6.4
Dubformer	8	10	7	16	13	6	10	6	12	11	8	14	7	6	10	8	14	13	9	5	7	9.5
Gemini 3.1 Pro	13	17	13	11	7	9	12	9	5	14	7	10	11	12	8	11	6	8	7	23	8	10.5
MakotoAI	10	16	16	6	6	17	4	16	14	6	11	4	16	9	15	7	12	9	12	17	15	11.3
DeepSeek V4 Pro	11	14	17	10	14	14	6	12	16	12	12	11	15	11	12	9	5	7	8	12	11	11.4
Qingqiu-MT-9B	16	15	11	18	19	12	18	7	11	16	10	17	13	14	4	15	10	11	14	14	12	13.2
Gemma 4 - 31B	14	18	9	22	24	11	21	13	10	21	15	12	12	13	13	14	19	19	21	6	14	15.3
Command A+	12	9	19	21	18	20	27	14	19	7	21	18	18	15	16	16	17	28	17	7	16	16.9
Mistral Medium 3.5	17	13	18	23	22	21	16	15	23	15	23	22	19	16	6	17	9	17	20	13	18	17.3
GPT OSS 120B	20	19	25	20	20	18	15	18	25	18	20	20	22	21	9	19	1	23	22	18	13	18.4
SalamandraTA	21	20	26	13	17	27	17	25	22	19	19	25	24	17	23	21	24	15	23	26	26	21.4
Qwen 3.6 - 27b	23	23	22	25	23	25	14	22	24	25	24	23	25	20	19	25	20	21	24	24	20	22.4
Gemma 4 - 4B	25	22	21	30	28	19	26	23	18	28	22	19	23	23	24	20	25	24	29	19	21	23.3
Ministral 3 - 14B	22	21	24	26	25	26	23	27	27	24	27	27	26	22	22	24	16	20	25	27	25	24.1
Tower 9B	24	25	31	17	30	24	25	28	30	22	28	30	28	19	18	22	22	27	30	30	23	25.4
Qwen 3.5 - 9b	28	28	28	28	26	28	20	21	26	30	26	29	27	26	27	27	23	25	26	29	22	26.2
GPT OSS 20B	29	27	29	27	27	29	19	26	29	27	25	26	30	28	26	28	21	26	27	28	24	26.6
GLM 4 - 9B	31	29	30	31	29	31	24	24	31	29	29	28	31	29	29	30	27	22	31	31	27	28.7

Table 1: FACET system rankings across the 21 language pairs. Each cell is the rank of a system within that language pair (1 = best), darker is better, and – marks a pair the system did not enter. Systems below the rule are ordered by mean rank. *Human (postediting)* attains the best mean rank, while *Human (from scratch)* sits mid-table.

polished translation over first-draft translation and does not simply favor human text. Another is that a reference-free LLM judge is more tolerant of translations that stay close to the source structure, as post-edited MT does, and penalizes the freer choices of translation from scratch. The two readings cannot be separated without human judgments, and we leave the question to the official evaluation.

Effect of the consistency pass. Comparing the full FACET ranking with the two-pass ablation FACET–C, the Consistency pass changes the score of 2,726 of 26,885 items (10.1%), but the system ranking is almost untouched. No system moves by more than 0.2 in mean rank and no two systems swap order. The effect on error-free detection is smaller still. Only 0.7% of segments receive a different error-free label under FACET than under FACET–C. Part of this is structural, since Consistency errors are document-level and rare relative to segment-level Fluency and Accuracy errors. The pass also recovers fewer errors than its design allows. It is the only pass that takes a whole docu-

ment as input, and the model struggles to maintain a reliable account of recurring items across such long contexts. This shows in the low number of Consistency errors it recovers and in the share of flagged spans it cannot localize. The divergences it does catch are cross-segment errors that Fluency and Accuracy cannot see, such as one entity rendered under several names within a document. We take the weak effect as a limitation of long-context reliability in the current model rather than as evidence that document-level Consistency does not matter (Kim, 2025a).

A qualitative example. Figure 2 shows the three passes on one document. Fluency and Accuracy stay within a segment. Fluency flags target-language faults such as the gender agreement in *für das Telčer Standort* and the misplaced quotation marks, and Accuracy flags meaning errors against the source, among them the region name *Žďárský-Krkonoše*, where the source region *Žďársko* is rendered as a mountain range, and *Tradition der Pilger*, which turns a file-making tradition into a tra-

A full Czech–German document evaluated by FACET. Every detected error is highlighted: Fluency , Accuracy , Consistency .

Segment 1 — source. <p>Pilníkárna v Křižánkách na Žďársku se stala kulturní památkou</p>

Segment 1 — target. <p>Die Bürstenfabrik in Křižánky im Žďárský-Krkonose ist zu einem Kulturdenkmal geworden</p>

Segment 2 — source. <p>Pilníkárna v Křižánkách na Žďársku, která je součástí historického areálu Kupcova vodního mlýna a pily, se stala kulturní památkou. Navzdory úpravám je autentickým svědectvím zaniklého řemesla výroby železných pilníků. V tiskové zprávě to za telčské pracoviště Národního památkového ústavu uvedla Ilona Ampapová. Na Vysočině je téměř 3000 nemovitých kulturních památek, z toho 16 národních kulturních památek. Křižánky jsou velkou vesnickou památkovou rezervací. Památkově chráněná byla zatím jen pila a mlýn zmiňovaný poprvé v roce 1713. "Celý areál, včetně pilníkárn, je doložen již na mapě stabilního katastru z roku 1839," uvedla Ampapová. Pilníkárna vznikla v roce 1905 a její současná podoba odpovídá stavebním úpravám spojeným se zavedením pilníkářského řemesla. "Navzdory dlčím proměnám ve 20. století se zde dochovala řada autentických konstrukcí a prvků, které svědčí o původní funkci objektu a jeho historické hodnotě," uvedla Ampapová. Provoz pily i pilníkářská výroba v areálu skončila v roce 1947. Křižánky jsou největší vesnickou památkovou rezervací na Vysočině. V roce 2012 se staly Vesnicí roku Kraje Vysočina. Pilníkářskou tradici v obci připomíná od roku 2011 naučná stezka.</p>

Segment 2 — target. Die Schleiferei in Křižánky am Žďárský-Berg, die Teil des historischen Geländes des Kupcova-Wasser-Mühlen- und Sägewerks ist, wurde zu einem Kulturdenkmal. Trotz der Anpassungen ist sie ein authentisches Zeugnis des verschollenen Handwerks der Herstellung von Eisen-Schleifscheiben. In einer Pressemitteilung gab Ilona Ampapová für das Telčer Standort des Nationalen Denkmalinstituts dies an. In der Region Vysočina gibt es fast 3000 unbewegliche Kulturdenkmäler, davon 16 nationale Kulturdenkmäler. Křižánky sind ein großes Dorfdenkmalreservat. Bisher waren nur der Sägewerk und der Mühle, die erstmals im Jahr 1713 erwähnt wurden, unter Denkmalschutz gestellt. „Das gesamte Gelände, einschließlich der Sägewerkshalle, ist bereits auf der Karte des festen Katasters aus dem Jahr 1839 belegt“, erklärte Ampapová. Die Schleiferei wurde im Jahr 1905 gegründet, und ihre heutige Gestalt entspricht den baulichen Anpassungen, die mit der Einführung des Schleifereihandwerks verbunden waren. „Trotz der teilweisen Veränderungen im 20. Jahrhundert sind hier zahlreiche authentische Konstruktionen und Elemente erhalten geblieben, die von der ursprünglichen Funktion des Objekts und seinem historischen Wert zeugen“, erklärte Ampapová. Der Betrieb des Sägewerks sowie die Sägewerkzeugherstellung auf dem Gelände endeten im Jahr 1947. Křižánky sind die größte dörfliche Denkmalschutzzone in der Region Vysočina. Im Jahr 2012 wurden sie zum Dorf des Jahres der Region Vysočina. Seit 2011 erinnert ein Lehrpfad an die Tradition der Pilger.

Figure 2: A full document evaluated by FACET, with all detected errors highlighted. Fluency and Accuracy act within a segment, from the target alone and from the source–target pair. Consistency acts across the document. The facility named *Bürstenfabrik* in Segment 1 becomes *Schleiferei* in Segment 2, a divergence that no single-segment judgment can detect.

dition of pilgrims. The Consistency pass catches an error of a different kind. The same facility is called *Bürstenfabrik* (‘brush factory’) in the title segment and *Schleiferei* (‘grinding shop’) in the body segment, two names for one referent that are each locally acceptable and conflict only when the segments are read together. This is the error class the pass is designed for, and it cannot be recovered by scoring segments in isolation. The example also shows a miss. Both *Bürstenfabrik* and *Schleiferei* mistranslate the source *pilníkárna* (‘file workshop’), and the Accuracy pass flags neither, so the document is penalized for the divergence but not for the term itself.

6 Conclusion

We presented FACET, a reference-free translation quality evaluator built by decomposition rather than by a single prompt or a trained metric. Error types need different context. Accuracy is judged from the source and the target together, while Fluency and Consistency are judged from the target alone, at the sentence and the document scope respectively. A single fixed model is prompted three times, and its predictions are merged and scored, with no trained components anywhere in the pipeline.

Without gold labels, we characterized the predictions of FACET on their own terms. Its per-language-pair system rankings follow the expected order of the participating systems and place post-edited human translation at the top. Whether the mid-table position of translation from scratch reflects

quality or a preference for source-close output remains open until human judgments are released. The document-level Consistency pass changes a tenth of segment scores but leaves the system ranking almost unchanged, a result our paired submissions were designed to expose. We attribute this to the limited reliability with which current models track recurring items across long inputs, and to the mismatch between a document-level signal and segment-level scoring. Stronger long-context models, or a scoring scheme that credits document-level Consistency directly, would address these two limits.

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); the NRF grant funded by the MSIT (RS-2025-16071992). This work was supported by Korea Internet & Security Agency(KISA) grant funded by the Korea government(PIPC) (No.RS-2026-25526342, Development of Technologies for Preventing Sensitive Information Inference and Risk Assessment in Foundation Model Operations). This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026 (Project Name: Develop AI agent technology to connect knowledge through public cultural facility-based discussion and communication, Project Number: RS-2026-25520645).

References

- Everlyn Asiko Chimoto and Bruce A. Bassett. 2022. [COMET-QE and active learning for low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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.
- 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.
- 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.
- Ahrii Kim. 2025a. [Do llms truly benefit from longer context in automatic post-editing?](#)
- Ahrii Kim. 2025b. [RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023a. [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 and Christian Federmann. 2023b. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. [Multidimensional quality metrics: a flexible system for assessing translation quality](#). In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. [MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.
- Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. [Error analysis prompting enables human-like translation evaluation](#)

in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. [TransQuest: Translation quality estimation with cross-lingual transformers](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Weilu Xu, Yunzhi Shen, Xinye Wang, Ranfei Dang, and Shujian Huang. 2026. [Rubric-as-experts: Case-specific mqm rubrics for translation quality evaluation](#). *Preprint*, arXiv:2606.21559.

Appendix

A Prompts

The three passes use the system prompts below, with {src_lang} and {tgt_lang} filled per language pair. The severity rubric in Appendix A.4 is appended verbatim to all three, and each error object in the JSON output carries an additional field "severity": 1-4. The user message is the target segment for Fluency, the source and target segment wrapped in <source> and <target> tags for Accuracy, and the full target document for Consistency.

A.1 Fluency Pass

System prompt: Fluency

You are a strict proofreader of {tgt_lang}. Find ALL spans that must change to make the text correct, natural {tgt_lang}; scan to the end and report every one. There may be several, or none.

Judge it purely as {tgt_lang} writing, by the norms of {tgt_lang} itself and never another language's. Apply {tgt_lang}-specific phenomena a native writer observes – e.g. paired *¿* *!* in Spanish, mandatory kana/kanji and no capitalization in Japanese, sentence-final *。* in CJK, German noun capitalization, French spacing before *;* *:* *?* *!*, Arabic/Hebrew script and directionality.

Categories:

- spelling: misspelling, wrong or missing diacritic.
- grammar: ungrammatical construction, wrong inflection, tense, or particle.
- agreement: features that must agree do not (gender, number, case, verb).
- word_order: an order no fluent writer of {tgt_lang} would use.
- punctuation: wrong or missing punctuation by {tgt_lang} convention.
- orthography: wrong case, spacing, script, or number/date/time/measurement formatting by {tgt_lang} convention (decimal comma, thousands separator, date and time format, unit spacing).
- register: register or variety inconsistent within the text.
- term: a word that is not the established {tgt_lang} term. Do not judge whether a term fits a domain – that needs the source.

Flag a span only if you are certain it is wrong. If your reason would contain "might", "could", "awkward", "more natural", or "better", it is a preference, not an error – do not flag it. Report each error separately, even when several occur close together. Do not judge meaning, numbers, or dates.

Output – JSON: {"errors": [{"quote": str, "span": str, "reason": str, "category": str}]}

- quote: copy the error with enough surrounding text to locate it uniquely in the text – roughly 3-6 words, verbatim.
- span: the exact minimal substring within quote that is the error.
- reason: one short clause, in English.
- category: one of the eight above.
- If nothing must change, return {"errors": []}.

User message: Fluency

{tgt_seg}

A.2 Accuracy Pass

System prompt: Accuracy

Check whether the {tgt_lang} target conveys the meaning of the {src_lang} source. Find ALL spans where the target departs from the source in meaning; scan to the end. There may be several, or none.

Judge ONLY meaning. Grammar, spelling, punctuation, and style are checked separately – do not report them here.

Categories:

- mistranslation: the target says something different from the source. Judge the target word by what the source means HERE, not by whether the target reads plausibly on its own. Cognates and literal renderings that "read fine" (e.g. "flanco" (Spanish) for "flank") are the most-missed errors – check them.
- wrong_term: a term valid in general but wrong for what the source refers to.
- addition: content in the target absent from the source (target side).

- hallucination: unsupported invented text with no basis in the source (target side). This is always severity 4 (major).
 - omission: content in the source absent from the target (source side).
 - untranslated: source-language text left untranslated (target side).
 - source_issue: the source itself is defective (typo, garbled, truncated) and cannot be translated correctly (source side). Only when the source is at fault; a fine source with a wrong target is a mistranslation.
- Do not flag paraphrase, reordering, or idiom that preserves meaning, or a difference the target language forces (gender, formality, pronouns).
- Flag a span only if you are certain the meaning is wrong. If your reason would contain "might", "could", or "possibly", or if the target is a defensible rendering, it is not an error – do not flag it.
- Output – JSON: [{"side": "src"|"tgt", "quote": str, "span": str, "reason": str, "category": str}]}
- side: "src" for omission or source_issue; otherwise "tgt". The quote is found on that side's text.
 - quote: copy the error with enough surrounding text to locate it uniquely on its side – roughly 3-6 words, verbatim.
 - span: the exact minimal substring within quote that is the error.
 - reason: one short clause – what the source means vs. what the target says.
 - category: one of the seven above.
 - If accurate, return {"errors": []}.

User message: Accuracy

```
<source>
{src_seg}
</source>
<target>
{tgt_seg}
</target>
```

A.3 Consistency Pass

System prompt: Consistency

You are given a full {tgt_lang} document. Check whether entities and key terms are rendered CONSISTENTLY across the whole document. Judge only internal consistency, not correctness.

Step 1. List the recurring items that must stay consistent: named entities (people, places, organizations), domain terms, and fixed expressions. Include items referred to across sentences, whether by the same words or by later shortened or pronominal references.

Step 2. For each item appearing more than once, check across the document:

- reference consistency: a single referent must not appear under TWO OR MORE different names/terms that a reader would take as different things (e.g. "McConnell" in one place, "Mitchell" in another). One consistent form across the document is correct, not an error.
- agreement across references: when a later shortened or pronominal form refers back (e.g. a name later taken up as "el coche"), its gender/number and any agreeing words match that chosen form correctly. Apply only if {tgt_lang} marks such agreement.

Flag a span only when one of these breaks. Only flag when you are certain both refer to the identical entity.

Do NOT flag:

- a normal shortened or pronominal reference itself – that is cohesion; flag only when its agreement breaks grammatically.
- an item that appears only once.
- anything about whether a rendering is correct – that needs the source.
- an item rendered the SAME way every time – identical renderings are consistency, never a conflict. Only flag when renderings actually DIFFER.

Output – JSON:

```
{"tracked": [{"referent": str, "renderings": [str, ...]}],
 "errors": [{"referent": str, "quote": str, "span": str, "reason": str, "category":
 "inconsistency"}]}
```

- tracked: EVERY recurring entity or term you checked from Step 1, each with ALL its renderings across the document as they literally appear – even when they are perfectly consistent. This is your audit trail; do not omit items.
- errors: only referents whose "renderings" list in tracked contains two or more genuinely different forms. If a tracked item has a single rendering, it CANNOT be an error.
 - referent: the tracked item this conflict belongs to (same string as in tracked).

- quote: copy the conflicting rendering with enough surrounding text to locate it uniquely in the document – roughly 3-6 words, verbatim.
- span: the exact minimal substring within quote that is the error.
- reason: name the referent and what it conflicts with.
- If nothing recurs, both lists are empty.

User message: Consistency

{tgt_seg}

A.4 Shared Severity Instruction

The following rubric is appended to all three system prompts, and each error object in the JSON output carries an additional field "severity": 1-4. The cut point that maps the scale onto Minor and Major is applied in post-processing, not in the prompt (Section 3).

Appended to all three system prompts

For each error, rate its severity on this 1-4 scale:
Scale 1: Slight wording change with no impact on clarity or intent.
Scale 2: Alters wording but message and intent remain mostly clear.
Scale 3: Noticeably distorts meaning or intent; the reader is misled or confused.
Scale 4: Substantial distortion; translation is unfaithful and potentially misleading.