

Findings of the WMT26 Shared Task on Automated Translation Quality Evaluation Systems: Compact Open Models are Competitive Quality Evaluators

Alon Lavie Carnegie Mellon University	Greg Hanneman Unaffiliated	Stefano Perrella Cohere	Shuoyang Ding NVIDIA
Eleftherios Avramidis DFKI	Lorenzo Proietti Sapienza University of Rome	Chi-kiu Lo 羅致翹 National Research Council Canada	Ammon Shurtz Brigham Young University
Chrysoula Zerva NTUA	Archchana Sindhujan University of Surrey	Vilém Zouhar ETH Zurich	Diptesh Kanojia University of Surrey
Frédéric Blain Tilburg University	Brian Thompson NVIDIA	Giorgos Filandrianos NTUA	
Orfeas Menis Mastromichalakis Instituto de Telecomunicações	Tom Kocmi Cohere	Pranav Gupta Cisco	

Abstract

We present the findings of the WMT26 Shared Task on Automated Translation Quality Evaluation Systems, continuing last year’s unification of the earlier separate WMT Metrics and Quality Estimation shared tasks. This year we evaluated three complementary views of segment-level translation quality on a common test set: fine-grained error-span detection, continuous quality-score prediction, and a new task on identifying error-free translations. Submissions to upstream tasks were also converted automatically to downstream predictions. The evaluation covered 21 translation directions using human cESA judgments from the WMT26 General Machine Translation task, with optional reference translations generated as either native, post-edited, or pseudo-references, depending on the translation direction. Official evaluation data was complemented by five submitted challenge sets. Across all three primary tasks, unsupervised LLM-as-a-judge approaches outperformed all traditional and supervised metrics, while open-weight models such as Gemma 4 were found to be competitive with proprietary frontier models. Reference-free LLM judges are highly competitive, and the benefit of adding a reference largely depends on its provenance and on the evaluator model family.¹

1 Introduction

The WMT26 Shared Task on Automated Translation Quality Evaluation Systems evaluates the

performance of state-of-the-art automated systems that assess the quality of text-to-text translations. It builds on many previous years of WMT shared tasks in both MT metrics and quality estimation, which were unified for the first time into a single track last year (Lavie et al., 2025). Automated translation quality evaluation systems play a critical role in the research, development, and deployment of machine translation systems, including multi-lingual LLMs. They are also critical components in automated translation workflows for large-scale commercial translation use cases.

The shared task consists of three primary subtasks and one additional subtask shown in Table 1. The primary subtasks are: (1) segment-level error detection and span annotation, (2) segment-level quality score prediction, and (3) detection of error-free segments. A common curated evaluation data set was provided for these three tasks, consisting of material obtained from the WMT26 Shared Task on General Machine Translation. It was augmented by a number of “challenge sets” developed by the task organizers and members of the research community. The fourth subtask solicited the submission of these challenge sets.

A central goal of this year’s shared task was to study these three primary evaluation formulations jointly rather than as independent problems. We therefore use a common evaluation set across Tasks 1–3 and organize them as a conceptual pipeline: fine-grained error annotations are also converted into segment-level quality scores, and both error annotations and quality scores are in turn also converted into binary error-free judgments. This de-

¹Web: www2.statmt.org/wmt26/mteval-task.html
Email: wmt-qe-metrics-organizers@googlegroups.com

	Source: Non parliamo italiano. Translation: I don't speak Spanish!
Task 1: error span prediction	→ I don't speak Spanish!
Task 2: score prediction	→ 25%
Task 3: error-free detection	→ False
Task 4: challenge sets	→ Source: ..., Translation: ...

Table 1: Illustration of the four subtasks: span-level error detection, segment-level quality score prediction, error-free detection, and challenge sets.

sign allows us to examine not only which systems perform best within each task, but also whether current automated evaluators can capture translation quality consistently across these different levels of granularity.

Below are some of the key details and changes implemented for this year’s shared task:

- **Subtasks:** We continue last year’s subtasks on segment-level quality score prediction, segment-level error-span/severity annotation, and the solicitation of challenge sets. The subtask on quality-informed error correction is dropped; it is replaced by a new subtask on detecting error-free segments.
- **Cross-Subtask Participation:** To broaden participation and facilitate a more comprehensive analysis of current MT evaluation capabilities, we released a single test set to be used in common across the three primary subtasks. We also defined mechanisms to automatically convert submissions from one subtask to another.
- **Language Pairs:** Following last year’s philosophy, we attempted to offer our shared task in as many of the language pairs supported by the WMT26 General Machine Translation task as possible. We ultimately released test data covering 21 of the 23 language pairs: see Table 2. Several of these pairs are new compared to 2025. To help guide participants for whom evaluating the full set proved computationally infeasible, we identified a priority subset of eight language pairs.
- **Additional Resources:** Preparation of this year’s test set brought together different types of references, different translation instructions per domain, and new guidelines for scoring translation quality. We made these resources available to participants.
- **Human Evaluation:** Human evaluation was car-

ried out by the WMT26 General Machine Translation task (Kocmi et al., 2026a). This year it used the Contrastive Error Span Annotation (cESA) protocol² (Zouhar et al., 2026a) with a variable number of human annotations per MT system, depending on the system’s performance (Zouhar et al., 2026b).

- **Meta-Evaluation:** We evaluate auto-evaluators using meta-metrics that differ substantially from those used in previous years. For segment-level error detection and span annotation, we adopt the MPP (“Match with partial overlap and partial credit”) F -score (Perrella et al., 2026); see Section 5.1. In segment-level quality score prediction, we newly use “group by screen” Pairwise Difference Pearson (DiIanni and Deutsch, 2025) at the segment level and maintain “group by source” Soft Pairwise Accuracy (Thompson et al., 2024) at the system level; see Section 6.1. Our new task on detection of error-free segments is meta-evaluated with the Matthews Correlation Coefficient (Matthews, 1975); see Section 7.1.

- **Code and Data Release:** The evaluation set, raw data for each subtask, and code to generate the meta-analyses in this paper are released publicly at <https://github.com/wmt-conference/wmt26-mteval>.

Our main findings are:

- **Unsupervised LLM-as-a-judge systems lead across all three primary tasks.** In both fine-grained error detection and quality-score prediction, the highest-ranking systems are dominated by prompted LLM evaluators without task-specific fine-tuning, while conventional lexical and supervised neural metrics rank substantially lower (Section 5.2 and Section 6.2). LLM-based systems similarly dominate detection of error-free segments (Section 7.2).
- **Evaluation performance depends strongly on the underlying LLM, but no single model family dominates every setting.** Gemini-based evaluators are particularly strong for error-span detection and quality scoring, while GPT-based systems lead the direct error-free classification submissions. Open-weight Gemma 4 systems are largely competitive with the strongest proprietary evaluators across tasks (Sections 5.2–7.2).
- **Strong LLM evaluators can operate effectively**

²github.com/zouharvi/cESA

without references, and the value of a reference depends on its provenance and quality.

Native human references consistently improve the Task 1 LLM baselines we analyze, whereas post-edited and pseudo-references sometimes reduce performance; corresponding reference-based variants generally improve quality-score prediction in Task 2 (Sections 5.2.1 and 6.3.2).

- **Reasoning can improve evaluation quality, but its effect is model-dependent and appears to saturate.** Reasoning substantially improves Qwen-based evaluators while providing smaller gains for Gemma, and increasing GPT-5.6 Sol from medium to xHigh reasoning yields essentially unchanged Task 1 accuracy (Section 5.2.2).
- **Task-specific optimization is not always necessary for error-free detection.** Strong Task 1 and Task 2 systems that were converted automatically to binary predictions are competitive with direct Task 3 submissions, and frequently fall within the same statistical significance clusters (Section 7.3.1).
- **Robust multilingual evaluation remains challenging.** Rankings vary across translation directions and can change substantially for lower-resource languages and when evaluation is expanded beyond the eight priority directions (Sections 6.2 and 7.3.2; Appendix B).
- **Challenge sets uncover gaps.** The challenge sets described in Section 8 expose pervasive complementary failure modes, including local metric disagreement, error-correction insensitivity, gender bias, Unicode robustness, and omission/addition asymmetries, that are largely hidden and not captured by aggregate benchmark performance.

The remainder of the paper is organized as follows. Section 2 defines the four subtasks and the conversion mechanisms connecting the three primary tasks. Section 3 describes the shared evaluation data and human annotations, and Section 4 summarizes participant systems and organizer baselines. Sections 5, 6, and 7 present the results and analyses for error-span detection, quality-score prediction, and error-free classification, respectively; Section 8 describes the submitted challenge sets. Section 9 summarizes our conclusions and directions for future work.

2 Tasks

We organized three primary subtasks that address segment-level translation quality from three perspectives: (1) detection of fine-grained error spans and error severities within the segment, (2) prediction of the segment’s overall quality score, and (3) classification of the segment as error-free or not. A fourth subtask solicited the submission of challenge sets that test the three judgments of quality under particular scenarios or in particular targeted ways, as defined by the challenge set designer in each case.

Newly this year, we ordered the three main subtasks as a conceptual pipeline and defined automatic mechanisms to convert submissions received by one subtask to the subtask(s) downstream of it.

Each subtask, as well as the automatic conversion mechanisms between subtasks, is described below.

2.1 Task 1: Segment-Level Error Detection and Span Annotation

Given a source text, its translation, and optionally a reference translation, the goal of Task 1 is to detect translation errors and identify the precise location of each error within the translation along with its severity. For each translation, participants are asked to predict error spans as character offsets (i.e., start and end character indices) as well as an error severity (`Major` or `Minor`) for each identified error. Ground-truth offsets and error severity are obtained from the cESA human annotations (Zouhar et al., 2026a) generated at the WMT General MT shared task (Kocmi et al., 2026a). Submissions are evaluated and ranked based on their ability to correctly detect translation errors and assess their severity, using the MPP meta-evaluation strategy (Perrella et al., 2026).³

2.2 Task 2: Segment-Level Quality Score Prediction

The goal of Task 2 is to predict a numerical quality score for each source–target segment pair in the evaluation set, with a reference translation optionally being provided. Participants this year were asked to predict the cESA score (Zouhar et al., 2026a), which ranges from 0 to 100 (higher is better). Submissions are evaluated and ranked based on their predictions’ correlations with human-annotated cESA scores — again generated by the

³Additional details are provided in Section 5.1.

General MT shared task — at both the segment and corpus levels.

We convert Task 1 annotations to Task 2 by assessing fixed penalties for each reported error, using an MQM-style weighting scheme. Beginning at a score of 100, we subtract 10 points per major error and 2.5 points per minor error.⁴ The minimum possible score is 0; any additional penalties that would further reduce the score are ignored. Translations annotated with a single major error spanning the entire segment are assigned a score of 0. Empty translations marked as a single major omission are also scored as 0.

2.3 Task 3: Detection of Error-Free Segments

Task 3 is a new subtask focused on a concrete application of translation quality evaluation systems: detection of error-free segments. While traditional quality estimation focuses on continuous scores (Task 2) or fine-grained error spans (Task 1), this task simplifies the objective to a binary classification: “Is this translation error-free?” The goal is to identify segments that can be published or consumed without further human intervention. Participants must predict a binary label for each source–target pair; references are not allowed. Submissions are evaluated and ranked based on their alignment with gold labels derived from human annotations, measured by the Matthews correlation coefficient.

We derive binary labels from a Task 1 submission via a simple test: segments where the Task 1 submission predicts no errors are labeled as error-free, while segments with one or more predicted error spans (including omissions) are labeled as containing errors. To convert Task 2 submissions, we asked Task 2 participants to provide a custom threshold for their models’ scores as well as an indication of whether the boundary is exclusive ($>$) or inclusive ($\geq$).⁵ Any segment where the model score passes the threshold is classified as error-free; all others are classified as containing errors.

2.4 Task 4: Challenge Sets

While the first three subtasks are focused on the development of stronger and better automated quality evaluation systems, the goal of this subtask is for

participants to create test sets with challenging evaluation examples that current evaluation systems fail to identify or score correctly. Challenge sets can relate to any (or all) of the other subtasks. In order to shed light on different evaluation perspectives, Task 4 takes place in a decentralized manner: contrary to the main subtasks, the test sets are not provided by the organizers but by different research teams, who are also responsible for analyzing and presenting the results.

This subtask is inspired by the *Build It or Break It: The Language Edition* shared task (Ettinger et al., 2017), which aimed at testing the generalizability of NLP systems beyond the distributions of their training data.

3 Evaluation Data

3.1 Data Sourcing and Preparation

Similar to previous years’ editions, the source sides, MT outputs, and reference texts for our shared task are mainly derived from the WMT26 Shared Task on General Machine Translation (Kocmi et al., 2026a).

The test data sources cover the news, social media, software, speech, educational, and fact-checking domains. In the General MT shared task, some of this test material was multimodal: the social-media content was provided with screenshot images, and the speech data was provided as video files with uncorrected automatic transcripts. However, for our shared task, we released as our primary version the text content of the social-media posts and a *corrected* transcript of the video files. (The images and videos remained available separately, as additional information that participants could use.)

Candidate translations were likewise sourced from General MT, in the form of *all* of the MT systems submitted to that shared task. This represents a difference from last year, when only a selection of the higher-performing translation systems was chosen. We also included among the candidate translations any available references, whether human- or automatically derived. Overall, our participants were therefore asked to judge the quality of up to 47 different translations per source segment, and in no case fewer than 23.

The language pairs in our official test set are summarized in Table 2. This year, the General MT shared task did not release native human references for all language pairs. Native references

⁴Note that the relative error weights in MQM are normally set to a 5-to-1 ratio. We decided to use a 4-to-1 ratio based on an analysis of last year’s data showing that it produced a slightly higher correlation with human-annotated ESA scores.

⁵If this information is not provided, we default to a threshold of ≥ 85 .

Code	Source	Target	Priority	Ref Type	# Src in Test	% Annot	# Pairs in Test	% Annot
ces-deu	Czech	German	✓	Human	438	100.0%	13,578	43.5%
ces-ukr	Czech	Ukrainian		None	415	100.0%	12,035	41.3%
ces-vie	Czech	Vietnamese	✓	Human	418	11.7%	12,958	10.6%
eng-arz	English	Arabic (Egypt)	✓	Post-edit	433	100.0%	12,990	100.0%
eng-bel	English	Belarusian		Pseudo	354	100.0%	10,974	56.1%
eng-ces	English	Czech		Post-edit	448	72.8%	13,440	33.4%
eng-deu	English	German		Post-edit	459	100.0%	13,770	49.7%
eng-ekk	English	Estonian	✓	Human	356	100.0%	10,680	67.7%
eng-hye	English	Armenian		Pseudo	364	100.0%	10,192	60.1%
eng-ind	English	Indonesian	✓	Pseudo	331	100.0%	10,261	100.0%
eng-isl	English	Icelandic	✓	Human	360	100.0%	11,160	61.1%
eng-kaz	English	Kazakh		Pseudo	438	100.0%	12,702	62.8%
eng-lij	English	Ligurian (Italy)		Human	411	100.0%	11,919	72.2%
eng-lld	English	Ladin (Italy)		Human	376	100.0%	10,152	51.2%
eng-rus	English	Russian		Pseudo	426	100.0%	13,206	84.3%
eng-sme	English	Northern Sámi	✓	Human	493	33.5%	13,311	19.0%
eng-tha	English	Thai		Pseudo	451	100.0%	13,981	100.0%
eng-ukr	English	Ukrainian		Pseudo	406	100.0%	12,180	89.2%
eng-zho_CN	English	Chinese (Simplified)		Pseudo	481	34.3%	13,949	17.1%
eng-zho_TW	English	Chinese (Traditional, Taiwan)		Pseudo	325	59.1%	9,100	31.6%
zho-jpn	Chinese	Japanese	✓	Post-edit	335	100.0%	10,385	100.0%

Table 2: Language pairs covered by the official test set. Eight of them were identified to participants as high-priority pairs; the nature of the available reference for each pair is as noted. A dynamically chosen portion of each language pair’s material was human-annotated. Columns show the number of source segments and source–target pairs in the test set, along with the fractions that were annotated.

were available in seven language pairs: ces-deu, ces-vie, eng-ekk, eng-isl, eng-lij, eng-lld, and eng-sme. In four pairs, the reference consisted of post-edited MT output: eng-arz, eng-ces, eng-deu, zho-jpn. The remaining 10 language pairs were originally *referenceless*. In nine of them (eng-bel, eng-hye, eng-ind, eng-kaz, eng-rus, eng-tha, eng-ukr, eng-zho_CN, eng-zho_TW), we created pseudo-references via a five-stage pipeline (Kanojia et al., 2026) in which seven LLMs and MT systems generate candidates under different prompt conditions, three reference-free QE metrics score them, and a per-document selector keeps one, with constrained LLM post-editing applied where the selection is weak. Selection was led by MetricX-24 and calibrated on common language pairs from the WMT25 human judgments, constrained by a structure gate and a language-identification penalty. We ran the final language pair, ces-ukr, as a completely reference-free scenario. The nature of the reference in each segment — human, post-edit, pseudo, or null — was labeled in the test set for participants to condition on if desired.

Two language pairs from General MT (eng-jpn and eng-kor) do not appear in our shared task. These were removed from our evaluation set when it became apparent that no gold-standard human quality annotations would be collected for eng-kor,

and the human annotations for eng-jpn were in doubt. Within the 21 language pairs that did comprise our official test set, we selected a reduced set of documents based on how likely we believed they were to be heavily covered by the General MT human annotations. (Annotations this year were carried out to varying extents for different qualities of MT systems; later documents are only likely to be covered for the best-performing systems.) This set consists of 160 documents for Czech-source language pairs and 120 documents for all others.

Our final official evaluation set thus comprised 252,919 segment pairs. With challenge sets included, the total size was 363,242 segment pairs.

Because of the significant size, we designated a subset of eight language pairs as high priority for participants who otherwise found the whole set computationally or economically infeasible. The priority pairs — ces-deu, ces-vie, eng-arz, eng-ekk, eng-ind, eng-isl, eng-sme, and zho-jpn — were chosen to cover a range of language families, resource levels, and reference types, while being biased towards settings where we expected the gold-standard human annotations to be more reliable.

Additional resources. In addition to identifying the reference type and the domain label of each segment, we released a separate set of domain-specific translation and evaluation instructions. The trans-

lation instructions consist of the same directions provided to the human translators, along with the consequent LLM prompts used for baselines. The evaluation instructions consist of the same information given to the human annotators in making gold-standard judgments of translation quality following the cESA annotation protocol (Zouhar et al., 2026a). Our intent is to provide participants of our shared task with a finer-grained definition of what a correct translation looks like in each domain and how to numerically score it on a 0–100 scale. Both sets of instructions are available [online per-domain](#), [online global](#), and by Kocmi et al. (2026a).

3.2 cESA Human Evaluation

As mentioned previously, human evaluation was carried out according to the Contrastive Error Span Annotation (cESA) protocol (Zouhar et al., 2026a), with a variable number of human annotations per MT system depending on the system’s performance (Zouhar et al., 2026b). The process runs independently for each language pair. After a burn-in period, during which all MT systems receive a fixed number of annotations, further source segments and candidate translations are selected dynamically until the annotation budget is exhausted. Each screen of the cESA interface presents the human annotator with three candidate translations for the same source segment.

The fraction of source segments and the fraction of source–target segment pairs from the test set that are covered by human evaluation thus vary from language to language. These statistics are illustrated in the rightmost four columns of Table 2. Many language pairs are well covered by the annotation campaign, but some are not. We received especially few annotations for ces-vie, eng-sme, and eng-zho_CN.

4 Participants

This section summarizes the systems submitted to our primary subtasks. The challenge set participants are detailed in Section 8.

4.1 External Submissions

Table 3 lists the submissions we received from the community for Tasks 1, 2, and 3. Some participants submitted a single model to multiple tasks; others submitted two variant models with the same name to the same task. These details are elided in the table, where we show only task-level participation per model name. With each submission,

we asked participants to also hand in a model card that included a free-text system description. These descriptions are given in Appendix A. Participants were finally asked to submit a separate system description paper explaining their full approach. See the appendix and the participants’ separate publications for more information on each submitted system.

4.2 Baselines

We use LLM-as-a-judge baselines alongside traditional automatic metrics, both neural and string-based, covering Tasks 1 and 2 directly; Task 3 predictions for every baseline are derived automatically via the conversion mechanisms of Section 2, with no special-casing for baselines. Some LLMs are prone to repeating error spans within a segment’s Task 1 output. We removed exact duplicates matching on start and end offset, keeping the major-severity annotation when the severities disagreed. We run all Task 1 baselines both with and without a reference translation, falling back to the reference-free configuration for language pairs that lack one. Exact prompts used for LLM-as-a-judge baselines are included in Appendix E.

We detail task-specific baselines below.

4.2.1 Task 1

xCOMET (Guerreiro et al., 2024) xCOMET XL (3.5B) and XXL (10.7B) are neural models trained to jointly predict MQM-style error spans and a sentence-level quality score for a translation, supporting source-only, source+reference, and reference-only modes. We run both XL and XXL in the source+reference and the source-only configuration with word-level predictions, giving four xCOMET baselines in total (XL/XXL × with/without reference).

GEMBA-ESA LLM-as-a-judge baselines For our remaining Task 1 baselines, we prompt a range of open- and closed-weight LLMs following a two-stage GEMBA-ESA protocol (Kocmi and Federmann, 2023b,a): in Stage 1, given the source, the hypothesis, and (where enabled) a reference translation, the model is prompted to enumerate error spans with severities; in Stage 2, the model is separately prompted, *conditioned* on its own Stage 1 output, to produce a final 0–100 quality score, which we also report as these systems’ Task 2 predictions (Section 4.2.2).

Stage 2 is capped at 64 output tokens for every

Submission	Organization	T1	T2	T3	Paper
AEGIS	IIT Patna; Shiv Nadar University	■			(Mishra et al., 2026)
Austriadactylus, a variant of Pterodactyl	1020 Group	■			(Garland and Berger, 2026)
Austriadraco, a variant of Pterodactyl	1020 Group			■	(Garland and Berger, 2026)
bytepop_pro10	8AI LLC + Rutgers University			■	(Xu, 2026)
bytepop_pro9	8AI LLC			■	(Xu, 2026)
Cohere CAT+	Cohere	■			(Kocmi et al., 2026b)
Cohere CAT+ ensemble	Cohere		■		(Kocmi et al., 2026b)
cuni-cat-v4	Charles University	■			(Hrabal and Bojar, 2026)
cuni-da-6d	Charles University		■		(Hrabal and Bojar, 2026)
cuni-v14	Charles University	■			(Hrabal and Bojar, 2026)
cuni-v14-regression	Charles University		■		(Hrabal and Bojar, 2026)
doc-FACET	Soongsil University	■	■		(Kim et al., 2026)
FACET	Soongsil University	■	■	■	(Kim et al., 2026)
First try	AIST			■	(Servajean et al., 2026)
fluency2-gemini35-esa	Algebras		■		(Mongush and Pukhov, 2026)
fluencyscore-multispan-v2	Algebras	■			(Mongush and Pukhov, 2026)
fluencyscore-multispan-v2.1	Algebras	■			(Mongush and Pukhov, 2026)
gemba-poly	Karlsruhe Institute of Technology		■		(Züfle et al., 2025)
gemma4-31b-v4ref-struct-best	Data Intelligence Lab, Hankuk University of Foreign Studies (HUFs)	■			(Jang et al., 2026)
Lexicala-QE-Ensemble (Majority-Vote)	Lexicala	■	■	■	(Siani, 2026)
Lexicala-QE-Ensemble (STAPLE)	Lexicala	■	■	■	(Siani, 2026)
MQM-LLM	AIST			■	(Servajean et al., 2026)
MQM-LLM Confidence Aware	AIST			■	(Servajean et al., 2026)
PhraseT1-01	Phrase a.s.	■			(Eckhardt et al., 2026)
PhraseT1-02	Phrase a.s. + CUNI	■			(Eckhardt et al., 2026; Hrabal and Bojar, 2026)
PhraseT2-01	Phrase a.s.		■		(Eckhardt et al., 2026)
PhraseT2-02	Phrase a.s. + CUNI		■		(Eckhardt et al., 2026; Hrabal and Bojar, 2026)
PhraseT3-01	Phrase a.s.			■	(Eckhardt et al., 2026)
PhraseT3-02	Phrase a.s. + CUNI			■	(Eckhardt et al., 2026; Hrabal and Bojar, 2026)
PragmaSpan	Brain Ventures	■			(Lee, 2026)
Pterodactyl	1020 Group		■		(Garland and Berger, 2026)
PTT-MULTITAN-v2	ALTAE, Université Paris Cité			■	(Zhu et al., 2026)
Qwen-based Omission	AIST	■			(Servajean et al., 2026)
Rule-based Omission	AIST	■			(Servajean et al., 2026)
QwenSpan	Soochow University	■			[Citation missing]
RAGthoven Error-Free	NaiveNeuron			■	[Citation missing]
Vertical_1	Université Grenoble Alpes, OVHai LLM		■		(Nakhlé et al., 2026)
Vertical_2	Université Grenoble Alpes, OVHai LLM		■		(Nakhlé et al., 2026)
Wukongopteris, a variant of Pterodactyl	1020 Group			■	(Garland and Berger, 2026)
XCOMOffset	University of Washington	■			(Pong, 2026)

Table 3: Participant submissions across our three primary subtasks. A ■ mark in column “T_n” means that the system was submitted directly to Task *n*; most (but not all) participants opted into automatic conversion to downstream tasks. “Paper” gives the citation for the corresponding system description.

baseline. Gemma 4 and Qwen 3.6 are served locally using vLLM (Kwon et al., 2023): we decode greedily (temperature 0) with a 512-token Stage 1 cap when reasoning is disabled, and sample at temperature 0.6, top-*p* 0.95, with an 8192-token Stage 1 cap when reasoning is enabled. Using the same token limits, Gemini 3.6 Flash is queried via its standard API at medium thinking level; GPT 5.6 via the OpenAI Batch API, at medium reasoning effort for Terra and at medium and xHigh for Sol. For the English-Armenian language pair, Qwen 3.6 and Gemma 4 token limits are multiplied $\times 4$ as many more tokens were produced consistently for segments in this language pair in order to have completed predictions. All reference-aware variants use the provided reference type (human, post-edited, or pseudo-reference) and adapt prompt wording accordingly.

The specific configurations we ran per model are given below:

- **QWEN 3.6 (Qwen Team, 2026)**: Qwen 3.6 35B-

A3B is an open-weight mixture-of-experts model with a toggleable reasoning mode. We run it with the GEMBA-ESA protocol above in all four combinations of reasoning enabled/disabled and reference used/not used.

- **GEMMA 4 (Team, 2026)** Gemma 4 31B-IT is an open-weight instruction-tuned model with a toggleable reasoning mode. As with Qwen 3.6, we run it in all four combinations of reasoning enabled/disabled and reference used/not used.
- **GEMINI 3.6 FLASH (Google DeepMind, 2026)** Gemini 3.6 Flash is a closed-weight model. We run it at medium reasoning effort and without a reference translation. We use the Google AI Studio API for querying.
- **GPT 5.6 (OpenAI, 2026)** We additionally run two closed-weight variants of GPT 5.6, Terra and Sol, both without a reference translation. GPT 5.6 Terra is run at medium reasoning effort; GPT

5.6 Sol is run at both `medium` and `xHigh` reasoning effort, giving three baselines in total. We limit all three baselines to the 8 priority language pairs due to budget constraints, with the `xHigh` reasoning GPT 5.6 Sol model able to complete 6 of the 8 pairs within timing constraints. The OpenAI Batch API is used for querying.

4.2.2 Task 2

We computed scores for several baseline systems in order to compare submissions against previous well-studied metrics.

Reused Task 1 baselines The `xCOMET` and `GEMBA-ESA` LLM baselines described in Section 4.2.1 natively produce a Task 2 quality score as part of the same prediction (for `xCOMET`) or the same two-stage prompting pipeline (for the LLM baselines). We reuse those predictions directly.

SacreBLEU baselines We used the following metrics from SacreBLEU (Post, 2018):

- **BLEU (Papineni et al., 2002)** is based on the precision of n -grams between the MT output and its reference, weighted by a brevity penalty. We used the SacreBLEU command line with default arguments with `-sl` argument to obtain segment-level BLEU.⁶
- **SPBLEU (NLLB-Team et al., 2022)** is the BLEU score computed with subword tokenization by the standardized FLORES-200 SentencePiece model. We used the SacreBLEU command line with `-sl` argument to obtain segment-level SPBLEU.⁷
- **CHRF (Popović, 2015)** uses character n -grams instead of word n -grams to compare the MT output with the reference. We used the SacreBLEU command line with default arguments and the `-sl` argument for segment-level CHRF.⁸

BERTSCORE (Zhang et al., 2020) leverages contextual embeddings from pre-trained transformers to create soft alignments between words in hypothesis and reference segments using cosine similarity. Based on the alignment matrix, BERTSCORE returns a precision, recall, and F1 score. We used F1 without TF-IDF weighting.

⁶nrefs:1lcase:mixedltok:13alsmooth:explversion:2.6.0. For into-Chinese, into-Japanese, and into-Korean language pairs, we used tok:zh, tok:ja-mecab, and tok:ko-mecab as the tokenizer, respectively.

⁷nrefs:1lcase:mixedltok:flores200lsmooth:explversion:2.6.0

⁸chrF2lnrefs:1lcase:mixedlnc:6lnw:0lspace:nolversion:2.6.0

COMET-22 (Rei et al., 2022a) is a learned metric fine-tuned using direct assessments from previous WMT translation shared tasks. This metric relies on sentence embeddings from the source, translation, and reference to produce a final score. We used the default model `wmt22-comet-da` provided in version 2.0.2 of the Unbabel/COMET framework. This model employs XLM-RoBERTa large as its backbone model and is trained on data from the 2017 to 2019 WMT shared tasks, in combination with the MLQE-PE corpus (Fomicheva et al., 2022).

COMETKIWI (Rei et al., 2022b) is a reference-free learned metric that functions similarly to BLEURT, but instead of encoding the translation along with its reference, it uses the source. We used the `wmt22-cometkiwi-da` model, which was a top-performing reference-free metric from the WMT 2022 shared task. This metric is fine-tuned on the same data as `wmt22-comet-da` using version 2.0.2 of the Unbabel/COMET framework.

YISI-1 (Lo, 2019) is an MT evaluation metric that measures the semantic similarity between a machine translation and human references by aggregating the IDF-weighted lexical semantic similarities based on the contextual embeddings extracted from pre-trained language models (BERT-base-chinese for evaluating Chinese and XLM-RoBERTa for evaluating other target languages in this shared task).

4.2.3 Task 3

For all Task 3 baselines, we use auto-conversion from both Task 1 and Task 2 baselines described above in Sections 4.2.1 and 4.2.2.

5 Task 1 Results and Analysis

5.1 Meta-Evaluation

Task 1 employs the MPP (“match with partial overlap and partial credit”) meta-evaluation measure recommended by Perrella et al. (2026). Given sets of hypothesis and ground-truth error spans, MPP establishes a one-to-one matching between them. It then computes per-span precision and recall, based on the proportion of overlapping characters between each hypothesis span and its matched ground-truth span. Sample-level (or corpus-level, depending on the desired granularity) precision and recall are obtained by averaging the per-span precision and recall values, respectively, and the final

F -score is computed as the harmonic mean of the resulting average precision and recall.

For each translation direction, we compute a single MPP F -score by micro-averaging across data samples. We do not report macro-averaged results across data samples, as this approach was found to be biased toward auto-evaluators that return fewer errors because of the sparsity of errors in MT error detection (Perrella et al., 2026). Finally, the MPP precision, recall, and F -score values obtained for each translation direction are macro-averaged to derive summary statistics for the entire test set.

5.1.1 Severity weighting

In previous years, span-level meta-evaluation measures were implemented with a 0.5 severity penalty to discount the contribution of errors whose predicted severity did not match the ground-truth severity (Blain et al., 2023; Zerva et al., 2024; Lavie et al., 2025). However, as noted by Lyu et al. (2026) and Perrella et al. (2026), this approach only penalizes severity mismatches and does not account for the fact that predicting a `Minor` error where no ground-truth error is present is less severe than predicting a `Major` error at the same location. More generally, `Major` errors should have a greater impact on the final F -score than `Minor` errors.

Thus, rather than applying a simple severity penalty, this year we adopt the severity weighting strategy proposed by Aur et al. (2026), whereby `Minor` and `Major` errors contribute to the final F -score in proportion to their respective weights, while severity mismatches remain penalized. Specifically, we assign weights of 0.25 and 1 to `Minor` and `Major` errors, respectively,⁹ such that each error span’s contribution to the final F -score is proportional to its severity weight.¹⁰

5.1.2 Implementation details

We compute MPP using the implementation provided in the `span-mt-metrics-eval` GitHub repository, with the following evaluation configuration:

```
from span_mt_metrics_eval import compute

predictions = [...]
references = [...]

result = compute(
```

```
    predictions,
    references,
    measure="MPP",
    matching="one_to_one",
    matching_algorithm="greedy",
    averaging="micro",
    severity_weights={
        "minor": 0.25,
        "major": 1.00,
    },
)
```

Reweighting multiply annotated translations.

The dynamic cESA annotation protocol employed in this year’s General MT shared task may result in the same translation being evaluated multiple times if it appears in more than one Pearmut screen.¹¹ To make use of all available annotations, we replicate each auto-evaluator assessment once for each ground-truth annotation associated with the same translation. However, without further adjustment, this would cause some assessments to weigh more heavily in the final precision, recall, and F -score simply because the corresponding translation was evaluated multiple times. We address this issue by reweighting the contribution of each assessment using the `segment_weights` input parameter of the `span_mt_metrics_eval.compute` function. Specifically, each replicated assessment is assigned a weight of $\frac{1}{N}$, where N is the number of ground-truth annotations associated with the corresponding translation. In this way, each translation contributes equally to the final precision, recall, and F -score, while all available human annotations are retained.

5.2 Main Results

Table 4 reports the ranking of all auto-evaluators, including participant submissions and organizer baselines, across the eight priority translation directions. Auto-evaluators that did not cover all eight translation directions are excluded from this ranking; their performance is reported only in the per-translation-direction results in Appendix B.1.

Unsupervised LLM-based auto-evaluators are state of the art

The top of the ranking is occupied entirely by unsupervised, LLM-based approaches to error detection, whereas the xCOMET baselines rank at the very bottom. This contrasts with last year’s results, when xCOMET models were still competitive with LLM-based approaches.

⁹As mentioned in Section 2.2, we derive these weights from an analysis of last year’s data, which showed that this weighting scheme yielded a slightly higher correlation with ESA scores.

¹⁰We refer the reader to Perrella et al. (2026) for the formal definition of MPP and to Aur et al. (2026) for the formal definition of the severity weighting strategy.

¹¹In cESA, a Pearmut screen consists of a source document and a set of three MT systems that are evaluated simultaneously in a contrastive manner.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	39.21	32.95	34.07
2	2	PhraseT1-01*	Gemini 3.5 Flash	34.73	36.34	33.47
3	2	fluencyscore-multispan-v2*	Gemini 3.5 Flash	40.39	31.39	33.46
4	2	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	40.27	31.44	33.46
5	3	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	33.79	36.04	32.85
6	4	Gemma 4 - thinking	Gemma 4 31B IT	30.89	39.03	32.58
7	5	Gemma 4	Gemma 4 31B IT	27.05	44.71	32.08
8	5	Gemma 4*	Gemma 4 31B IT	29.20	39.61	32.02
9	5	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	28.45	40.65	31.98
10	6	Gemma 4 - thinking*	Gemma 4 31B IT	32.16	35.03	31.72
11	6	GPT-5.6 Terra*	GPT-5.6 Terra	31.37	34.79	31.50
12	7	cuni-cat-v4*	Gemma 4 31B	33.58	32.80	31.35
13	8	cuni-v14*	Gemma 4 31B	33.10	32.76	31.13
14	9	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	26.08	38.21	29.77
15	10	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	23.79	43.01	29.38
16	10	PhraseT1-02*	Ensemble ^{E1}	22.74	44.85	29.17
17	11	Cohere CAT+*	Command A+	27.04	30.78	26.93
18	12	Qwen 3.6*	Qwen3.6-35B-A3B	17.50	46.44	24.62
19	13	FACET*	GPT-4.1 Mini (2025-04-14)	24.22	26.89	23.79
20	14	doc-FACET*	GPT-4.1 Mini (2025-04-14)	24.19	26.93	23.78
21	14	Qwen 3.6	Qwen3.6-35B-A3B	15.74	52.11	23.46
22	15	xCOMET XXL [†]	xCOMET-XXL	13.89	31.28	17.85
23	16	xCOMET XL [†]	xCOMET-XL	11.31	37.94	16.74
24	17	xCOMET XXL ^{*†}	xCOMET-XXL	11.98	26.42	13.93
25	17	Qwen-based Omission ^{*†}	Qwen3-32B	17.81	13.41	13.81
26	18	Rule-based Omission ^{*†}	Qwen3-32B	17.80	13.35	13.78
27	18	xCOMET XL ^{*†}	xCOMET-XL	9.33	35.95	13.77

Table 4: Task 1 ranking for the set of eight priority translation directions. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]. Similar to previous years (Freitag et al., 2023, 2024; Lavie et al., 2025), statistical significance uses the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021a), with clusters computed greedily ($p \leq 0.05$, unadjusted, two-sided, 10000 draws).

We also note that most submissions are unsupervised, while the only two fine-tuned LLM-based auto-evaluators (Qwen-based Omission and Rule-based Omission) also rank near the bottom. A promising direction for future research in MT error detection is therefore to investigate whether, and under what conditions, fine-tuning remains beneficial relative to unsupervised LLM-as-a-Judge approaches.

Auto-evaluators based on the same underlying model occupy similar ranks

Submissions and baselines group together according to their underlying model. The top four positions are all occupied by Gemini-based auto-evaluators. The only Gemini-based system that ranks lower is PhraseT1-02, which is an ensemble of auto-evaluators based on three models: Gemini 3.5 Flash, Gemma 4 12B, and Gemma 4 31B. By contrast, PhraseT1-01 – which is based solely on Gemini 3.5 Flash – ranks third, suggesting that ensembling Gemini predictions with those of other models reduced per-

formance relative to using Gemini alone. However, this result may be attributable to the specific ensembling strategy rather than to the models themselves. In particular, all error annotations produced by the individual models were retained in the ensemble prediction, leading to higher recall at the cost of lower precision.

Ranks 5 through 13 are predominantly occupied by auto-evaluators based on Gemma 4 31B, with the exceptions of ranks 9 and 11, occupied by GPT 5.6 Sol and GPT 5.6 Terra, respectively. Interestingly, all reference-based Gemma 4 auto-evaluators rank higher than their reference-free counterparts. We examine the performance differences between reference-based and reference-free auto-evaluators in greater detail in Section 5.2.1.

Ranks 14 through 21 comprise auto-evaluators based on Command A+, Qwen 3.6 35B A3B, and GPT 4.1 Mini, together with the PhraseT1-02 ensemble. The F -score range across these ranks is wider, spanning from 29.77 to 23.46, with the lowest score obtained by the reference-based Qwen 3.6

Auto-evaluator	Ref-free	Ref-based	$\Delta F1$
Gemma 4 [#]	31.72	32.58	+0.86
Qwen 3.6 [#]	29.77	29.38	-0.39
Gemma 4	32.02	32.08	+0.06
Qwen 3.6	24.62	23.46	-1.16
xCOMET XL	13.77	16.74	+2.97
xCOMET XXL	13.93	17.85	+3.92
	Non-thinking	Thinking	$\Delta F1$
Gemma 4	32.08	32.58	+0.49
Qwen 3.6	23.46	29.38	+5.91
Gemma 4 [*]	32.02	31.72	-0.31
Qwen 3.6 [*]	24.62	29.77	+5.15

Table 5: Effect of reference use and thinking in task 1. [#] marks thinking auto-evaluators. ^{*} marks reference-free auto-evaluators.

variant with reasoning disabled.¹²

Finally, ranks 22 through 27 are occupied entirely by fine-tuned auto-evaluators, including the widely used encoder-based xCOMET models.

Auto-evaluators are recall-oriented. Across most submissions and baselines, recall is higher than precision. This tendency is particularly pronounced among reference-based evaluators, which we find to predict more errors than their reference-free counterparts (see Section 5.2.1). Interestingly, three of the four Gemini-based auto-evaluators achieve higher precision than recall, suggesting that Gemini is more conservative than other models in detecting translation errors. Similarly, Qwen-based Omission and Rule-based Omission—two fine-tuned, Qwen-based auto-evaluators—also achieve higher precision than recall, possibly as a consequence of their fine-tuning setup.

5.2.1 Are references still needed?

To measure the effect of reference translations, we analyze auto-evaluators that were run in both reference-based and reference-free settings. Specifically, we compute the MPP F -score under both conditions and report the resulting $\Delta F1$ scores in Table 5. As expected, references are highly beneficial for xCOMET models, which were pre-trained less extensively than modern LLM-based approaches and are therefore likely to have more

¹²We caution that the results for the Qwen 3.6 baselines should be interpreted carefully, as we observed many duplicated errors and several cases of endless repetition, particularly in variants run with reasoning enabled. It is unclear whether this issue resulted from suboptimal generation parameters—which were nevertheless kept identical across all baselines, as reported in Section 4.2—or from a failure mode of the model itself.

$\Delta F1$ = with reference – without reference

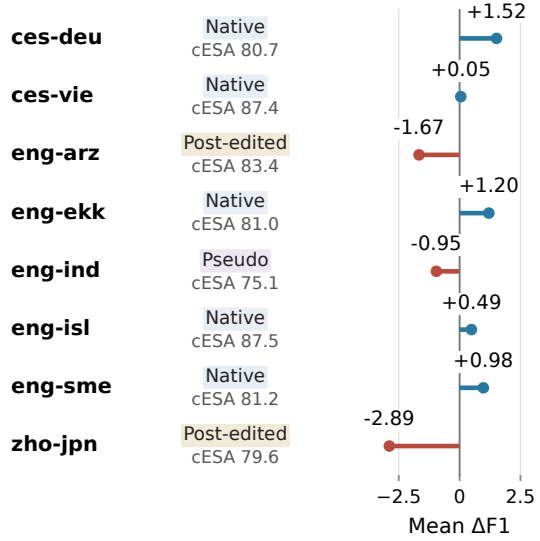


Figure 1: Breakdown of the effect of reference use on MPP F -score across translation directions. Average scores only include Gemma 4 and Qwen 3.6 baselines.

limited multilingual capabilities. References may thus simplify the task and allow these models to perform effectively across a wider range of translation directions. By contrast, the effect of reference use on the LLM-as-a-Judge baselines is mixed: references mildly improve performance for Gemma 4, while slightly degrading it for Qwen 3.6. We investigate this pattern further by measuring the effect of references for each translation direction, averaged across the Gemma 4 and Qwen 3.6 models, and report the results in Figure 1.

An interesting pattern emerges: reference use improves performance, albeit sometimes only marginally, in all settings where the provided reference was produced by human translators from scratch—i.e., Native in the figure. By contrast, Post-edited and Pseudo-references appear to be detrimental to evaluation performance, with average reference-free Gemma 4 and Qwen 3.6 scores higher than the average of their reference-based counterparts.

One possible explanation is reference quality, if Native references were systematically of higher quality than Post-edited and Pseudo references. However, the average cESA scores of the references across translation directions, reported in Figure 1, do not support this as the sole explanation. For example, the Post-edited references for en→arz achieve a higher average cESA score than many Native references. We therefore hy-

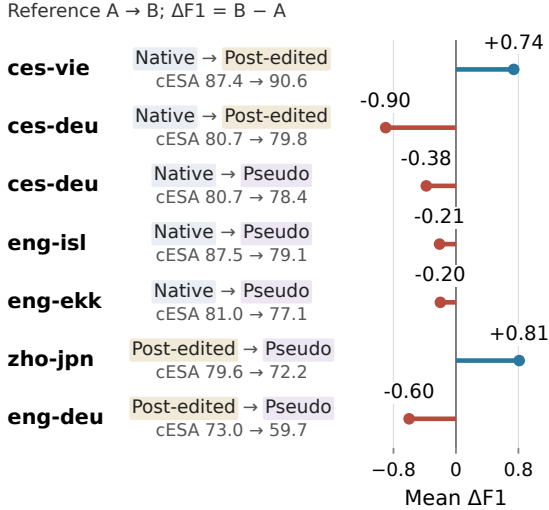


Figure 2: Effect of reference type and quality across translation directions. Average scores only include Gemma 4 and Qwen 3.6 baselines.

pothesize that the observed effect may also be related to the greater similarity of `Post-edited` and `Pseudo` references to other system translations, which could bias reference-based evaluation in favor of systems producing similar outputs.

The impact of reference type on reference-based evaluation We investigate the extent to which reference type and quality affect the performance of reference-based evaluation. To this end, we compare the performance of reference-based auto-evaluators across different reference types and report the results in Figure 2. Specifically, we average the $\Delta MPP F$ -score across five reference-based auto-evaluators, including both submissions and baselines, for three pairwise comparisons: (1) `Native` vs. `Pseudo`, (2) `Native` vs. `Post-edited`, and (3) `Post-edited` vs. `Pseudo`.

Across translation directions covered by multiple reference types, `Native` references appear to provide the strongest signal: in only one case (ces $\rightarrow$ vie) do `Native` references yield a lower average MPP F -score than `Post-edited` or `Pseudo` references. Indeed, in this case, the `Post-edited` reference received a higher cESA score than the `Native` reference (90.6 vs. 87.4). More generally, in six of the seven translation directions considered, the reference type associated with the higher MPP F -score is also the one assigned a higher cESA score by human annotators. This suggests that reference quality may be an important factor underlying the performance differ-

ences observed in Figure 2. The only exception is zho $\rightarrow$ jpn, where a lower-quality `Pseudo` reference leads to a higher MPP F -score than a higher-quality `Post-edited` reference.

Reference-based auto-evaluators find more, shorter errors We investigate whether the use of reference translations makes auto-evaluators more severe or more lenient. To this end, we measure the effect of references on the number of predicted errors, average error length, and average error severity, and report the results in Table 6.

The pattern is consistent across both auto-evaluators and across reasoning settings: when reference translations are available, models tend to identify more translation errors, while the detected error spans are, on average, shorter than those produced by reference-free auto-evaluators. We hypothesize that access to reference translations makes certain errors easier to detect and localize more precisely, resulting in the observed increase in the number of predicted errors and decrease in average error span length. Finally, we observe a slightly lower average severity weight, suggesting that reference-based LLM-as-a-Judge auto-evaluators identify a greater proportion of `Minor` errors relative to `Major` errors than their reference-free counterparts.

5.2.2 What is the impact of thinking?

Similar to the previous analysis of reference-based versus reference-free evaluation, this section examines the impact of thinking on MPP F -score and on the number, length, and severity of detected errors. The analysis is restricted to the Gemma 4 and Qwen 3.6 baselines, as these are the only auto-evaluators that were run both with and without thinking.

As shown in Table 5, thinking has different effects on the Qwen 3.6 and Gemma 4 baselines, substantially improving performance for the former while having little to no effect on the latter. The statistics in Table 6 provide further insight into how thinking affects these auto-evaluators’ predictions. Here, the pattern is consistent across models: auto-evaluators with thinking detect fewer and shorter errors than their non-thinking counterparts. In addition, thinking reduces the average severity weight, indicating that auto-evaluators with thinking identify a higher proportion of `Minor` errors relative to `Major` errors. All of these effects are substantially stronger for Qwen 3.6 than for Gemma 4, which may explain why the corresponding differences in

Auto-evaluator	Errors per prediction			Mean span length			Mean severity weight		
	Ref-free	Ref-based	Δ	Ref-free	Ref-based	Δ	Ref-free	Ref-based	Δ
Gemma 4 [#]	2.40	3.13	+0.73	21.20	18.03	-3.16	0.57	0.52	-0.05
Qwen 3.6 [#]	3.11	4.10	+0.99	21.43	20.14	-1.29	0.55	0.50	-0.04
Gemma 4	2.61	3.65	+1.04	21.69	20.32	-1.36	0.61	0.53	-0.07
Qwen 3.6	3.93	5.23	+1.29	27.43	26.49	-0.94	0.67	0.64	-0.03
	Non-thinking	Thinking	Δ	Non-thinking	Thinking	Δ	Non-thinking	Thinking	Δ
Gemma 4	3.71	3.18	-0.53	19.69	17.50	-2.19	0.53	0.51	-0.02
Qwen 3.6	5.36	4.15	-1.21	25.84	19.88	-5.95	0.63	0.50	-0.13
Gemma 4 [*]	2.67	2.44	-0.24	21.25	20.27	-0.98	0.60	0.57	-0.04
Qwen 3.6 [*]	3.97	3.15	-0.82	27.34	20.95	-6.39	0.67	0.54	-0.13

Table 6: Effect of reference use and thinking on predicted errors in task 1. [#] marks thinking auto-evaluators. ^{*} marks reference-free auto-evaluators. Errors per prediction include omissions; mean span length is computed over non-omission spans; and mean severity weight includes omissions.

MPP F -score are considerably larger for Qwen 3.6.

Overall, references and thinking have effects in the same direction with respect to error span length and severity weight, as both reduce average span length and mean error severity. By contrast, they have opposite effects on the number of predicted errors: reference-based auto-evaluators identify more errors, whereas auto-evaluators with thinking identify fewer.

GPT Sol Medium vs. xHigh In the context of our analysis of thinking, we also investigate the effect of increasing the thinking level from Medium to xHigh for one of our baselines, namely GPT 5.6 Sol. However, we were unable to run GPT 5.6 Sol with xHigh thinking on the full set of eight priority translation directions, and some API queries exceeded our maximum allowed number of completion tokens. To ensure a fair comparison between the two settings, we therefore restrict the analysis to the subset of translations for which both variants completed their reasoning and returned valid outputs.

The results in Table 7 show that increasing thinking effort has little to no impact on MPP F -score. GPT 5.6 Sol with xHigh thinking identifies slightly more errors, but these errors are also slightly shorter on average. As a result, precision, recall, and F -score remain nearly unchanged. These findings suggest that lower thinking effort may achieve comparable MT evaluation accuracy while reducing token usage. However, the scope of this analysis is limited, as it considers only a single model at two thinking levels. We therefore encourage future work to investigate this question more systematically.

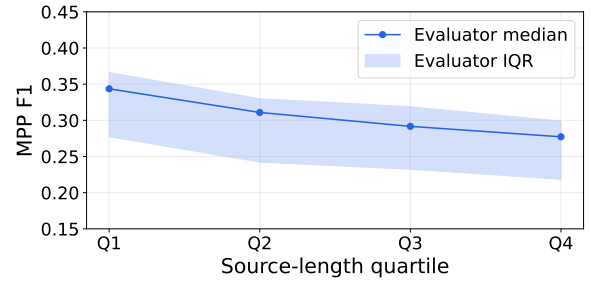


Figure 3: Average auto-evaluator MPP F -score at varying source lengths.

5.2.3 Does evaluation accuracy depend on text length?

Here, we investigate the impact of source length on evaluation accuracy. Specifically, for each of the eight priority translation directions, we divide the data into four quartiles based on source segment length and compute the median MPP F -score for each quartile across all auto-evaluators and translation directions. We measure source length as the number of source tokens returned by the XLM-RoBERTa tokenizer (Conneau et al., 2020).

The results are presented in Figure 3, which shows that MPP F -score decreases as source length increases. This suggests that auto-evaluators are less accurate at detecting errors in translations of longer source segments.

6 Task 2 Results and Analysis

6.1 Meta-Evaluation

Task 2 evaluates whether an auto-evaluator’s score judgments align with human cESA quality scores. We use complementary meta-metrics at two granularities. At the segment level, the relevant question is whether the evaluator assigns an appropriate score difference to competing translations of the

Auto-evaluator	P	R	F1	Errors per prediction	Mean span length	Mean severity weight
GPT-5.6 Sol - medium	23.08	38.78	28.94	2.89	21.41	0.40
GPT-5.6 Sol - xhigh	23.42	37.79	28.92	3.07	19.68	0.38
Δ (xhigh – medium)	+0.34	−0.99	−0.02	+0.18	−1.73	−0.02

Table 7: GPT-5.6 Sol thinking-effort comparison.

same source item. At the system level, the question is whether aggregating those judgments leads to the same conclusions as humans when comparing MT systems. Accordingly, the final Task 2 score is the equally weighted mean of a segment-level and a system-level meta-metric, after first averaging each component equally across the eight priority translation directions. This design also reflects a broader shift in MT evaluation toward comparative judgments: recent human evaluation protocols increasingly assess translations side by side, improving agreement and sensitivity to subtle quality differences (Song et al., 2025; Zouhar et al., 2026a). A similar direction is emerging for automatic evaluation, with recent methods increasingly evaluating candidate translations comparatively rather than in isolation (Moosa et al., 2024; Züfle et al., 2025; Proietti et al., 2026).

Screen-based PDP. For the segment-level component, we use Pairwise Difference Pearson (DiIanni and Deutsch, 2025, PDP). PDP concatenates score differences between translations of the same source item and computes one Pearson correlation between the automatic and human difference vectors. Unlike the segment-level acc_{eq}^* meta-metric used for the score prediction task in the last three editions of the shared task (Deutsch et al., 2023; Freitag et al., 2023, 2024; Lavie et al., 2025), PDP preserves the magnitude of a predicted preference without computing a tie-calibration threshold from the test data. The combination of differences for the same source into a single correlation further avoids noisy segment-wise Pearson correlations based on few observations, while excluding comparisons between unrelated source items (DiIanni and Deutsch, 2025), which could favor auto-evaluators that rely on patterns spuriously correlated with translation quality (Perrella et al., 2024b).

Importantly, PDP has also been shown to better reward human evaluators than acc_{eq}^* , which can disadvantage their more discrete score distributions and the resulting ties relative to the continuous scores produced by many automatic metrics (Proietti et al., 2025a; DiIanni and Deutsch, 2025). This

property may also be relevant for LLM-as-a-judge methods: at WMT25, these methods were among the strongest auto-evaluators at the system level, but they performed substantially worse at the segment level under acc_{eq}^* . Since LLM judges can also produce comparatively discrete score distributions and exact ties, PDP may provide a more suitable segment-level meta-evaluation of such evaluators.

The WMT26 General MT human evaluation uses contrastive error span annotation (cESA): a Pearmut screen presents three translations of one source item simultaneously, and annotators assign each translation both error spans and a holistic score (Zouhar and Kocmi, 2026; Zouhar et al., 2026a; Kocmi et al., 2026a). We therefore use screen-based PDP (“group by screen”). For every pair of translations shown together in a screen, we add both ordered score differences to the PDP vectors. Thus, a three-translation screen contributes six ordered differences. If the same item and system pair occurs on multiple screens, we first average its available human scores within that co-screen comparison, so it contributes once. This construction relies exclusively on the direct contrastive comparisons elicited by cESA, which are the most reliable form of human evidence available in this setting (Zouhar et al., 2026a).

Source-based SPA. For the system-level component, we use Soft Pairwise Accuracy (Thompson et al., 2024, SPA). For every pair of MT systems, SPA compares the p -values obtained from permutation tests on the human and automatic segment scores, and awards soft credit according to their agreement. It thus evaluates not only whether an auto-evaluator prefers the same system as humans, but also whether it reflects the uncertainty of that comparison. We use 1,000 permutation resamples for every SPA calculation, consistent with the original SPA setup and the last two editions of the shared task (Thompson et al., 2024; Freitag et al., 2024; Lavie et al., 2025), and give equal weight to every pair of MT systems.

We use source-based SPA (“group by source”): for a system pair, we retain every source item on

which both systems have a human score, regardless of whether their translations appeared together on a screen. When a translation was assessed on multiple screens, its human score is the arithmetic mean of its retained assessments. We prefer it to screen-based SPA because it retains substantially more common source items per pair, avoiding the highly sparse and heterogeneous co-screen support induced by the contrastive assignment. Appendix C analyzes both grouping decisions (i.e., group-by-screen and group-by-source) on the eight primary directions and compares their resulting rankings, as the choice of grouping can fundamentally alter the evaluation question being posed and, consequently, the interpretation of meta-evaluation results (Perrera et al., 2024a; Proietti et al., 2025b).

6.2 Main Results

Table 8 reports the Task 2 ranking across the eight priority translation directions. It includes only direct Task 2 participant submissions and baselines that produced scores for every translation with human scores in all eight directions. The 40 eligible auto-evaluators comprise 18 participant submissions and 22 baselines. Systems with incomplete coverage appear only in the per-direction results in Appendix B.2. PDP and SPA are averaged equally across directions, and *Avg Corr* is their arithmetic mean.

Unsupervised LLM auto-evaluators are state-of-the-art The first 16 positions are occupied by auto-evaluators based on unsupervised LLM-as-a-Judge approaches, and none is fine-tuned, as marked in Table 8. The highest-ranked system not based on an LLM is chrF at rank 17, followed by COMET22 at rank 19. Thus, the aggregate point estimates favor LLM judging without task-specific fine-tuning over the conventional lexical and supervised metrics included here.

Auto-evaluators based on the same underlying model occupy similar rank The ranking also groups systems by their underlying model. Gemini 3.6 Flash ranks first. The GPT-5.4 and GPT-5.6 systems occupy ranks 2, 5, 8, and 9; the Gemma 4 31B systems rank 6, 7, 10, 12, 15, and 16; and the Qwen 3.6 systems rank 11, 13, 14, and 18. The pattern is not determined by the base model alone. Reference use, reasoning, prompting, and ensembling still change results within each family.

Results vary by translation direction Across the priority directions, the winner is not always the same: GPT-5.6 Sol Medium ranks first for Czech to German and English to Indonesian; GPT-5.6 Terra for Czech to Vietnamese; Gemini 3.6 Flash for English to Arabic; byteword_pro10 for English to Estonian and Icelandic; byteword_pro9 for English to Northern Sámi; and the reference-based Gemma 4 auto-evaluator with reasoning ranks first for Chinese to Japanese. At the same time, each direction’s ranking is moderately to strongly associated with the eight-direction ranking (Spearman $\rho = 0.838\text{--}0.952$). The direction tables show larger changes outside the priority set: spBLEU ranks first for English to Ligurian and English to Latin, despite ranking 22nd in Table 8. This result reinforces evidence that multilingual automatic evaluators, including reference-free LLM systems, remain often unreliable in low-resource settings (Sindhuja et al., 2025; Li et al., 2025).

6.3 Analysis

6.3.1 Automatic Task 1-to-Task 2 conversions

We evaluate automatic Task 1-to-Task 2 conversions separately from the direct Task 2 ranking. For each predicted Task 1 assessment, the conversion starts at 100 and subtracts 10 points for each Major error and 2.5 points for each Minor error, with a lower bound of zero (Section 2.2). Thus, these values are determined directly by the submitted error spans rather than by an independently predicted holistic score.

Table 9 reports the conversion track on the eight priority directions only. It includes the 36 systems with complete converted predictions on every human-supported translation: 12 Task 1 participant submissions and 24 organizer baseline variants.

For the LLM-as-a-Judge baselines, we additionally test the effect of removing repeated predicted spans. Repeated spans have a limited effect on a directly submitted Task 2 score, which is elicited as a holistic judgment in the second stage of the GEMBA-ESA prompt. They can have a much larger effect after automatic conversion because every retained span directly incurs an error penalty. We compare the original baseline outputs with variants deduplicated by identical start offset, end offset, severity, and category. Positive differences below refer to the deduplicated score minus the original score.

The effect is concentrated in Qwen 3.6. Its

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.736	0.924	0.830
2	1	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.735	0.920	0.828
3	2	bytepop_pro10	Ensemble ^{E2}	0.733	0.918	0.826
4	3	bytepop_pro9	Ensemble ^{E2}	0.730	0.917	0.823
5	3	GPT-5.6 Terra*	GPT-5.6 Terra	0.723	0.920	0.821
6	4	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.714	0.914	0.814
7	5	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.708	0.909	0.808
8	5	Pterodactyl*	GPT-5.4	0.707	0.910	0.808
9	5	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.705	0.911	0.808
10	6	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.691	0.908	0.800
11	6	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.682	0.917	0.800
12	7	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.683	0.905	0.794
13	8	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.669	0.904	0.786
14	9	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.617	0.911	0.764
15	9	cuni-da-6d*	Gemma 4 31B IT	0.637	0.887	0.762
16	10	cuni-v14-regression*	Gemma 4 31B IT	0.618	0.889	0.754
17	11	chrF	–	0.601	0.890	0.746
18	11	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.593	0.891	0.742
19	12	COMET22	XLM-RoBERTa-large	0.610	0.862	0.736
20	12	Cohere CAT+ ensemble*	Command A+	0.582	0.877	0.729
21	12	Vertical_2 [†]	Gemma 3 1B IT	0.582	0.877	0.729
22	13	spBLEU	–	0.566	0.885	0.725
23	14	AEGIS	XLM-RoBERTa-large; mT5-large	0.589	0.825	0.707
24	14	BLEU	–	0.539	0.873	0.706
25	15	Vertical_1 [†]	Gemma 3 1B IT	0.538	0.856	0.697
26	15	gemba-poly*	Llama 3.3 70B	0.510	0.884	0.697
27	16	xCOMET XXL [†]	xCOMET-XXL	0.540	0.822	0.681
28	16	MQM-LLM ^{*†}	Not specified (27B)	0.503	0.852	0.678
29	16	BERTScore	XLM-RoBERTa-large	0.496	0.858	0.677
30	17	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.493	0.852	0.672
31	17	xCOMET XL [†]	xCOMET-XL	0.518	0.804	0.661
32	17	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.464	0.850	0.657
33	18	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.402	0.883	0.642
34	18	PhraseT2-02*	Ensemble ^{E1}	0.432	0.849	0.641
35	19	YiSi-1	XLM-RoBERTa-large	0.378	0.866	0.622
36	20	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.363	0.840	0.601
37	20	FACET*	GPT-4.1 Mini (2025-04-14)	0.363	0.839	0.601
38	20	xCOMET XXL ^{*†}	xCOMET-XXL	0.416	0.757	0.586
39	21	COMETKiwi22*	XLM-RoBERTa-large	0.426	0.724	0.575
40	22	xCOMET XL ^{*†}	xCOMET-XL	0.398	0.719	0.558

Table 8: Task 2 ranking on the eight priority translation directions. PDP and SPA are equally averaged across directions; Avg Corr is their arithmetic mean. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	GPT-5.6 Terra [no dedup]*	GPT-5.6 Terra	0.557	0.900	0.729
2	2	GPT-5.6 Terra [dedup]*	GPT-5.6 Terra	0.555	0.900	0.728
3	3	Gemini-3.6-Flash [no dedup]*	Gemini 3.6 Flash	0.525	0.882	0.703
4	3	Gemini-3.6-Flash [dedup]*	Gemini 3.6 Flash	0.525	0.881	0.703
5	4	Gemma 4 Reference-based - Reasoning [no dedup]	Gemma 4 31B IT	0.498	0.883	0.690
6	4	Gemma 4 Reference-based - Reasoning [dedup]	Gemma 4 31B IT	0.497	0.883	0.690
7	4	cuni-cat-v4*	Gemma 4 31B IT	0.504	0.868	0.686
8	4	Gemma 4 - Reasoning [dedup]*	Gemma 4 31B IT	0.492	0.879	0.686
9	4	Gemma 4 Reference-based - No Reasoning [no dedup]	Gemma 4 31B IT	0.496	0.876	0.686
10	4	Gemma 4 - Reasoning [no dedup]*	Gemma 4 31B IT	0.492	0.879	0.686
11	5	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	0.491	0.876	0.684
12	5	Gemma 4 - No Reasoning [no dedup]*	Gemma 4 31B IT	0.498	0.869	0.684
13	5	Gemma 4 Reference-based - No Reasoning [dedup]	Gemma 4 31B IT	0.492	0.875	0.683
14	6	Gemma 4 - No Reasoning [dedup]*	Gemma 4 31B IT	0.495	0.869	0.682
15	6	cuni-v14*	Gemma 4 31B IT	0.495	0.867	0.681
16	7	Qwen 3.6 - Reasoning [dedup]*	Qwen3.6-35B-A3B	0.466	0.871	0.669
17	7	Qwen 3.6 Reference-based - Reasoning [dedup]	Qwen3.6-35B-A3B	0.447	0.878	0.663
18	8	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.455	0.862	0.659
19	8	FACET*	GPT-4.1 Mini (2025-04-14)	0.455	0.862	0.658
20	8	fluencyscore-multispan-v2*	Gemini 3.5 Flash	0.462	0.851	0.657
21	8	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	0.462	0.851	0.656
22	8	Cohere CAT+*	Command A+	0.439	0.864	0.652
23	8	PhraseT1-01*	Gemini 3.5 Flash	0.449	0.851	0.650
24	9	Qwen 3.6 - Reasoning [no dedup]*	Qwen3.6-35B-A3B	0.367	0.866	0.617
25	10	Qwen 3.6 - No Reasoning [no dedup]*	Qwen3.6-35B-A3B	0.333	0.823	0.578
26	11	Qwen 3.6 - No Reasoning [dedup]*	Qwen3.6-35B-A3B	0.330	0.820	0.575
27	11	PhraseT1-02*	Ensemble ^{E1}	0.334	0.802	0.568
28	11	Qwen 3.6 Reference-based - Reasoning [no dedup]	Qwen3.6-35B-A3B	0.281	0.843	0.562
29	12	Qwen 3.6 Reference-based - No Reasoning [dedup]	Qwen3.6-35B-A3B	0.266	0.819	0.542
30	13	Qwen 3.6 Reference-based - No Reasoning [no dedup]	Qwen3.6-35B-A3B	0.211	0.800	0.506
31	14	Qwen-based Omission ^{*†}	Qwen3-32B	0.154	0.785	0.469
32	14	Rule-based Omission ^{*†}	Qwen3-32B	0.153	0.784	0.469
33	14	xCOMET XXL [†]	xCOMET-XXL	0.202	0.709	0.456
34	15	xCOMET XXL ^{*†}	xCOMET-XXL	0.154	0.698	0.426
35	15	xCOMET XL [†]	xCOMET-XL	0.135	0.669	0.402
36	16	xCOMET XL ^{*†}	xCOMET-XL	0.069	0.621	0.345

Table 9: Task 1-to-Task 2 conversion ranking on the eight priority translation directions. Each score is the equal-direction mean; Avg Corr is the arithmetic mean of PDP and SPA. * marks reference-free systems. † marks fine-tuned systems. Light-grey rows are organizer baselines. [dedup] and [no dedup] identify the two Task 1 baseline snapshots. The four xCOMET pairs are identical before and after deduplication and are shown once. Ensemble E1 contains Gemini 3.5 Flash, Gemma 4 12B, and Gemma 4 31B. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

reference-based reasoning variant improves from 0.562 to 0.663. The reference-free reasoning and reference-based no-reasoning variants similarly gain +0.052 and +0.037, respectively. The largest absolute change is +0.101 for Qwen 3.6 reference-based reasoning, confirming that the baseline family most affected by deduplication is Qwen 3.6. The remaining variants move by at most 0.004. The four xCOMET pairs are identical before and after deduplication, and appear only once in Table 9.

Direct score prediction is substantially stronger than span-to-score conversion; A comparison of Table 9 with the direct Task 2 results in Table 8 reveals a consistent gap between holistic score prediction and scores reconstructed from Task 1 error annotations that are not due to duplicate spans. For example, Gemini 3.6 Flash obtains an Avg Corr of 0.830 as a direct Task 2 evaluator but 0.703 after converting its Task 1 spans; GPT-5.6 Terra drops from 0.821 to 0.729, and the reference-based reasoning variant of Gemma 4 drops from 0.814 to 0.690. The same pattern holds for the xCOMET baselines. Thus, while the Task 1 annotations contain information that can be converted into a meaningful overall quality score, they do not appear to constitute a sufficient representation of the holistic quality judgment. A fixed weighted count of detected Major and Minor errors clearly fails to capture information about the nature and cross-dependency of errors, as well as any quality signals not directly represented by its predicted spans. This contrasts with the Task 1/2-to-Task 3 conversions in Section 7.3.1, where converted predictions are often competitive with systems optimized directly for the downstream task.

6.3.2 Effects of reference translations

We examine reference use in three complementary ways. First, Table 59 in Appendix D compares strictly matched reference-free and reference-based baselines on the eight priority directions. Each pair uses the same underlying model and decoding setting. As in Task 1, references are most beneficial for xCOMET: the reference-based XL and XXL variants improve Avg Corr by +0.103 and +0.095, respectively. The gains are smaller but consistently positive for the Gemma 4 and Qwen 3.6 LLM-as-a-Judge pairs (+0.014 and +0.014 for the reasoning variants). Thus, the value of a reference depends strongly on the evaluator family.

We next hold the MT outputs and human evidence fixed while changing only the reference

supplied to reference-based auto-evaluators. The available challenge material does not span all eight priority directions: the main human versus PseudoRef comparison is available on 4 priority directions, and the from-scratch versus post-edited comparison on 2. For the former, the main human reference is from scratch for Czech to German, English to Estonian, and English to Icelandic, and post-edited for Chinese to Japanese. Tables 60 and 61, also in Appendix D, report these results. We exclude all designated and alternative reference candidates from both sides of each comparison in the test set as translations to score. We include only direct Task 2 systems with complete predictions on the matched challenge set samples.

Replacing the main human reference with PseudoRef lowers Avg Corr for 13 of 17 auto-evaluators. In contrast, from-scratch versus post-edited human references have less impactful changes: nine systems decrease and eight increase. The available evidence therefore suggests that replacing a human reference with PseudoRef can alter Task 2 performance, whereas the two human reference production methods are nearly interchangeable in this small matched sample.

Finally, Table 62 in Appendix D describes the submitted score scales themselves on human-supported translations, rather than meta-evaluation performance. This is the Task 2 analogue of Task 1’s analysis of whether references make an auto-evaluator more severe or lenient. Gemma 4 and the reasoning Qwen 3.6 baseline become more severe when supplied with a reference, with mean reference-based minus reference-free shifts of -2.4 and -2.8 points. In contrast, xCOMET XL and XXL shift upward by $+5.7$ and $+4.8$ points. These shifts should not be interpreted as quality gains or losses: they describe raw score behavior.

These experiments, along with Task 1 results, suggest that strong LLM evaluators are considerably less dependent on references than earlier learned metrics, while also exposing that different types of available references are not equally useful.

6.3.3 Reasoning-level impacts

The benefit of reasoning is strongly model-dependent. Table 8 allows controlled comparisons for Gemma 4 and Qwen 3.6, for which we ran otherwise matched variants with and without reasoning in both reference-based and reference-free settings. For Gemma 4, enabling reasoning produces only a small improvement: “Avg Corr” increases from

0.808 to 0.814 with a reference and from 0.794 to 0.800 without one. The effect is considerably larger for Qwen 3.6, increasing from 0.764 to 0.800 in the reference-based setting and from 0.742 to 0.786 in the reference-free setting.

For Qwen 3.6, most of the improvement comes from segment-level discrimination. With a reference, reasoning increases PDP from 0.617 to 0.682 while SPA increases only from 0.911 to 0.917. Without a reference, PDP increases from 0.593 to 0.669, compared with an SPA increase from 0.891 to 0.904. The corresponding changes for Gemma 4 are much smaller at both granularities. Thus, for Qwen, reasoning primarily improves the evaluator’s ability to distinguish quality differences among translations of individual source segments, while system-level rankings are already relatively stable without reasoning.

This pattern closely mirrors Task 1, where reasoning substantially improves Qwen 3.6 but has little effect on Gemma 4 (Section 5.2.2). Across the two tasks, the evidence therefore does not support reasoning as a uniformly beneficial property of an evaluator; rather, its value appears to depend strongly on the underlying model family.

6.3.4 LLM judges are strong at both levels of evaluation granularity

One motivation for adopting PDP as the segment-level meta-metric was the pronounced gap observed at WMT25 between the system- and segment-level performance of LLM-as-a-Judge evaluators. As shown by (Perrella et al., 2024b), acc_{eq}^* penalizes evaluators that produce discrete assessments (LLM judges or human annotators) compared to auto-evaluators that produce assessments on a continuous scale. Indeed, last year’s LLM judges were among the strongest system-level evaluators but performed substantially worse at the segment level. We do not observe the same pattern under this year’s Task 2 evaluation.

The strongest LLM judges rank highly on both components of the Task 2 meta-evaluation. Gemini 3.6 Flash and GPT-5.6 Sol obtain PDP correlations of 0.736 and 0.735, respectively, compared to 0.601 for chrF and 0.610 for COMET22. Their system-level SPA values are likewise strong at 0.924 and 0.920, although SPA scores are more compressed across evaluator families: chrF reaches 0.890 and COMET22 0.862. Thus, the advantage of the leading LLM judges is particularly pronounced at the segment level under PDP and is still evident after

aggregation to the MT-system level.

This result should not be interpreted as establishing that PDP alone explains the difference from WMT25: the test data, human evaluation protocol, evaluator models, and meta-evaluation measure all changed between editions. Nevertheless, it shows that poor segment-level performance is not an inherent property of LLM judges, and highlights that conclusions about automatic evaluators can depend on the choice of meta-evaluation methodology.

7 Task 3 Results and Analysis

7.1 Meta-Evaluation

Since Task 3 evaluates a binary label with potential label imbalance, we use the Matthews correlation coefficient (Matthews, 1975, MCC) as the main metric for this task. We also report precision, recall, and F_1 score to give a clearer picture of a system’s performance (with “error-free” as the positive label). For ranking and significance testing, we macro-average across translation directions.

For each metric pair, we run a permutation test (100,000 draws) on the difference in macro-average MCC, exchanging predictions jointly at the source-segment level.¹³ We then cluster greedily as in recent WMT shared tasks (Freitag et al., 2024; Lavie et al., 2025): descending the ranking, a metric joins the current cluster only if it does not differ significantly ($p < 0.05$, unadjusted two-sided) from every member, otherwise it starts a new one. Separate clusters therefore do not imply that every cross-cluster pair differs significantly. Further details are given in Appendix F.1.

7.2 Main Results

Table 10 reports rankings for the 18 direct Task 3 submissions and the converted organizer Task 1 baselines that cover all 8 priority language pairs. Auto-converted submissions, language pairs beyond the 8 priority ones, and Task 2 baseline conversions are discussed later in Section 7.3.

LLM-based submissions are state-of-the-art
As with Task 1 and Task 2, LLM-as-judge has become the de facto standard for error-free detection as well. In fact, among the systems included in Table 10, all but the two xCOMET baselines are LLM-based, and those two baselines are also the

¹³Following PERM-INPUTS (Deutsch et al., 2021b): one decision per segment, applied to every associated prediction, including repeats across language pairs. The 2,357 items form 1,392 such units.

Rank	Cluster	Submission	Underlying model	MCC $\times 100$	P(%)	R(%)	F1(%)
1	1	Wukongopterus	GPT-5.4	32.55	43.71	78.45	53.72
2	1	Austriadraco	GPT-5.4	31.63	47.70	63.43	52.23
3	1	Gemini 3.6 Flash	Gemini 3.6 Flash	30.98	53.15	47.98	48.20
4	2	GPT-5.6 Terra	GPT-5.6 Terra	29.38	52.45	46.02	45.16
5	2	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	28.42	55.89	37.62	41.39
6	2	Gemma 4 Reasoning	Gemma 4 31B IT	27.93	47.97	51.29	47.94
7	2	PhraseT3-01	Gemini 3.5 Flash	27.72	41.03	76.41	50.36
8	3	PhraseT3-02	Gemini 3.5 Flash	26.96	41.54	72.52	50.37
9	3	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	26.29	54.32	34.81	40.49
10	3	Gemma 4 No Reasoning	Gemma 4 31B IT	25.82	54.35	34.22	40.98
11	4	First try [871721]	Qwen3.6-27B	20.20	52.81	24.23	31.55
12	5	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	16.17	52.58	15.49	23.06
13	5	First try [874374]	Babel-83B-Chat	15.67	34.15	81.03	44.54
14	5	RAGthoven Error-Free	Ensemble ^{E1}	15.28	37.68	51.45	41.16
15	6	FACET [869556]	GPT-4.1 Mini (2025-04-14)	11.30	43.49	15.72	21.25
16	6	FACET [869557]	GPT-4.1 Mini (2025-04-14)	11.26	43.41	15.77	21.29
17	7	xCOMET XXL	xCOMET-XXL	6.62	36.86	32.98	28.64
18	8	xCOMET XL	xCOMET-XL	-0.98	29.15	10.92	13.00

Table 10: WMT26 Task 3 priority-direction ranking: direct Task 3 submissions and converted organizer Task 1 baselines covering all 8 priority language pairs (ces-deu, ces-vie, eng-arz, eng-ekk, eng-ind, eng-isl, eng-sme, zho-jpn), macro averaged over those pairs. MCC is the primary metric (adjusted with $\times 100$ to align with the scale for other metrics); precision (P), recall (R) and F1 take error-free as the positive class. Light-grey rows are organizer baselines. Omitted for not covering every pair: Lexicala-QE-Ensemble (Majority-Vote) (6 of 8), Lexicala-QE-Ensemble (STAPLE) (6 of 8) and PTT-MULTITAN-v2 (no priority pairs). Their results appear in the per-language tables. Cluster groups systems that are not significantly different, by a paired permutation test over source segments (100,000 permutations, two-sided, $\alpha = 0.05$) on the macro-averaged MCC, with WMT’s greedy clustering. Ensemble components named in the model cards are ^{E1} RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

lowest ranked, which further underscores the dominance of LLM-based systems. The per-language breakdown tables show that xCOMET-XXL still achieves mid-tier performance on language pairs covered by prior human evaluation data (e.g., eng-isl, eng-ukr, eng-zho), but generalizes poorly to low-resource languages and to language pairs that do not involve English.

Recall seems to be the overall bottleneck Looking at the precision and recall of the systems, many of them reach a similar precision level of 0.40–0.55. The range of recall, however, is much wider, which makes recall the more decisive of the two. This concurs with the observation in the Task 1 results (Section 5.2), where many LLMs err on the “cautious” side and tend to flag errors, with Gemini 3.5 Flash again being an outlier.

7.3 Analysis

7.3.1 Comparison with Converted Submissions from Task 1 and 2

For a more holistic view of how different approaches to Task 3 compare with each other, we conducted automatic conversions of Task 1 and

Task 2 submissions following the methodology outlined in Section 2.3. Although it was not required, all systems that did not opt out of automatic Task 2 to Task 3 conversion provided a threshold for such conversions, which we use throughout this section. The optimality of these thresholds is discussed in Section 7.3.3. Table 11 shows results including automatically converted systems on the same 8 priority language pairs as Table 10. This reveals a few further trends that were not apparent in Table 10.

Task-specific optimization for error-free detection does not clearly pay off

Once converted systems are included in the comparison, one natural question to ask is: do predictions from Task 3 participants significantly outperform binary labels derived from automatic conversions of Task 1 and Task 2 submissions? Table 11 gives a largely negative answer: while Wukongopterus remains the top-scoring Task 3 submission, its margin over all of its Task-2-converted siblings is too small to be considered significant. The same holds for other participants that submitted to Task 3 and at least one other track: the FACET (Soongsil University) Task 3 system does not meaningfully outper-

Rank	Cluster	Submission	Underlying model	Track	MCC _{×100}	P(%)	R(%)	F1(%)
1	1	Wukongopterus	GPT-5.4	T3	32.55	43.71	78.45	53.72
2	1	Austriadactylus	Ensemble ^{E1}	T2	32.01	46.13	69.99	52.37
3	1	Pterodactyl	GPT-5.4	T2	31.90	46.59	67.16	52.75
4	1	Austriadraco	GPT-5.4	T3	31.63	47.70	63.43	52.23
5	1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	30.98	53.15	47.98	48.20
6	2	cuni-v14	Gemma 4 31B	T1	30.20	47.18	60.99	51.53
7	2	cuni-cat-v4	Gemma 4 31B	T1	30.18	47.44	60.02	51.37
8	2	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	29.38	52.45	46.02	45.16
9	2	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	28.77	51.22	46.35	46.22
10	2	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	28.62	51.81	44.65	45.58
11	2	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	28.42	55.89	37.62	41.39
12	3	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	27.93	47.97	51.29	47.94
13	3	PhraseT3-01	Gemini 3.5 Flash	T3	27.72	41.03	76.41	50.36
14	3	PhraseT3-02	Gemini 3.5 Flash	T3	26.96	41.54	72.52	50.37
15	3	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	26.89	52.38	40.79	43.44
16	4	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	26.29	54.32	34.81	40.49
17	4	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	25.82	54.35	34.22	40.98
18	5	Cohere CAT+ ensemble	Command A+	T2	21.02	62.23	17.30	25.38
19	5	First try [871721]	Qwen3.6-27B	T3	20.20	52.81	24.23	31.55
20	5	gempa-poly	Llama 3.3 70B	T2	19.85	46.28	35.85	38.23
21	5	Cohere CAT+	Command A+	T1	19.75	46.77	32.48	35.63
22	6	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	16.17	52.58	15.49	23.06
23	6	First try [874374]	Babel-83B-Chat	T3	15.67	34.15	81.03	44.54
24	6	RAGthoven Error-Free	Ensemble ^{E2}	T3	15.28	37.68	51.45	41.16
25	6	FACET	GPT-4.1 Mini (2025-04-14)	T1	13.31	47.44	15.51	21.50
26	6	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	13.31	47.46	15.48	21.47
27	7	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	11.30	43.49	15.72	21.25
27	7	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	11.30	43.49	15.72	21.25
29	7	FACET	GPT-4.1 Mini (2025-04-14)	T2	11.26	43.41	15.77	21.29
29	7	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	11.26	43.41	15.77	21.29
31	8	Qwen-based Omission	Qwen3-32B	T1	7.43	33.89	45.00	35.14
32	9	Rule-based Omission	Qwen3-32B	T1	7.34	33.82	45.06	35.13
33	9	xCOMET XXL	xCOMET-XXL	T1-B	6.62	36.86	32.98	28.64
34	10	xCOMET XL	xCOMET-XL	T1-B	-0.98	29.15	10.92	13.00

Table 11: WMT26 Task 3 priority-direction ranking across all tracks: every system covering the same eight priority language pairs as Table 10, macro averaged over those pairs, with converted Task 1 and Task 2 submissions added. Metrics and row shading are as in Table 10. Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Conversions whose source system used the reference translations are excluded, Task 3 being a reference-free task. Omitted for not covering every pair: Lexicala-QE-Ensemble (Majority-Vote) (6 of 8; T1, T2, T3), Lexicala-QE-Ensemble (STAPLE) (6 of 8; T1, T2, T3), PTT-MULTITAN-v2 (no priority pairs; T3) and XCOMOffset (4 of 8; T1). Their results appear in the per-language tables. Cluster is defined as in Table 10. Ensemble components named in the model cards are ^{E1} Austriadactylus: GPT-5.4; GPT-5.6 Terra; ^{E2} RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

Rank	Submission	Underlying model	ces-deu	ces-vie	eng-arz	eng-ekk	eng-ind	eng-isl	eng-sme	zho-jpn	Avg.
1	Wukongopterus	GPT-5.4	23.92	27.71	32.21	38.55	31.91	42.61	27.94	35.59	32.55
2	Austriadrac0	GPT-5.4	23.60	24.87	26.75	41.99	29.96	42.81	28.96	34.12	31.63
3	Gemini 3.6 Flash	Gemini 3.6 Flash	21.70	24.24	24.38	47.72	32.96	38.77	34.43	23.67	30.98
4	GPT-5.6 Terra	GPT-5.6 Terra	20.33	21.55	22.83	46.96	31.29	37.85	31.49	22.76	29.38
5	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	22.82	18.11	20.77	47.17	30.51	34.62	34.62	18.72	28.42
6	Gemma 4 Reasoning	Gemma 4 31B IT	21.79	28.10	25.47	37.44	33.39	34.57	13.76	28.89	27.93
7	PhraseT3-01	Gemini 3.5 Flash	22.68	24.28	19.93	34.91	32.51	36.32	21.05	30.08	27.72
8	PhraseT3-02	Gemini 3.5 Flash	23.71	23.20	19.44	37.97	29.27	36.11	19.31	26.68	26.96
9	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	21.97	24.15	21.36	41.49	27.25	27.99	28.44	17.64	26.29
10	Gemma 4 No Reasoning	Gemma 4 31B IT	18.93	22.48	16.20	39.99	26.31	33.47	21.92	27.25	25.82
11	First try [871721]	Qwen3.6-27B	12.67	13.39	13.28	31.59	22.42	29.29	19.55	19.40	20.20
12	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	10.66	5.48	14.66	37.87	21.47	26.63	3.81	8.74	16.17
13	First try [874374]	Babel-83B-Chat	10.89	16.55	13.28	14.83	23.17	17.49	10.74	18.41	15.67
14	RAGthoven Error-Free	Ensemble ^{E1}	16.55	7.41	12.96	7.84	14.71	23.12	12.81	26.88	15.28
15	FACET [869556]	GPT-4.1 Mini (2025-04-14)	7.14	-0.39	9.62	27.64	12.12	12.19	11.41	10.67	11.30
16	FACET [869557]	GPT-4.1 Mini (2025-04-14)	7.24	-0.39	9.39	27.57	12.12	12.12	11.41	10.63	11.26
17	xCOMET XXL	xCOMET-XXL	-7.66	-6.81	11.29	20.67	23.07	21.84	-5.23	-4.24	6.62
18	xCOMET XL	xCOMET-XL	-1.40	-6.53	-13.85	10.56	9.89	8.69	-7.07	-8.13	-0.98
	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble ^{E2}	19.45	—	22.22	42.35	30.33	—	31.44	22.64	—
	Lexicala-QE-Ensemble (STAPLE)	Ensemble ^{E3}	17.95	—	22.12	42.42	29.44	—	33.30	19.43	—

Table 12: WMT26 Task 3 priority-direction ranking, per-language-pair MCC ($\times 100$). Avg. is the macro average that determines the ranking; the best value per language pair is in bold. Light-grey rows are organizer baselines. Below the rule are the systems that covered only some of the priority pairs: they are unranked and have no average, since it would be taken over a different set of pairs, but their per-pair scores are shown where they competed. Ensemble components named in the model cards are ^{E1} RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro; ^{E2} Lexicala-QE-Ensemble (Majority-Vote): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; ^{E3} Lexicala-QE-Ensemble (STAPLE): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus.

form the automatic conversions of its Task 1 and Task 2 systems. The Task 3 system from Lexicala (not shown here because of its language coverage) meaningfully outperforms the automatic conversions in only one language pair (eng-ekk). The only exception is the First try system from AIST, which significantly outperforms the automatic conversions of its Task 1 systems. It is worth pointing out that their Task 3 systems use very different base models from their Task 1 systems, so the two sets of systems may have been developed separately.

Source system matters more than source task for conversion Comparing the performance of binary labels converted from Task 1 with those converted from Task 2, we see no clear trend as to which is the better conversion path. Instead, we notice that conversions from the same system tend to cluster together in the rankings. This corroborates a trend noted in Section 5.2: systems based on the same underlying model tend to occupy similar ranks. The choice of the base system to convert from, and in particular the choice of a strong underlying model, therefore appears to be the most consequential decision here.

Compact, open-weight models show potential

In last year’s results, large closed-source language models such as Gemini 2.5 Pro and GPT-4.1 dominated the top of the LLM-as-judge ranking. For

Task 3, one departure from that result is that some systems based on relatively small open-weight models, such as cuni-v14 and cuni-cat-v4, are competitive with closed-source frontier models served behind APIs (e.g., GPT-5.6 Terra, GPT-5.6 Sol, Gemini 3.6 Flash, Gemini 3.5 Flash). Since neither of these two systems relies on fine-tuning, we believe there is further potential in specializing small open-weight models to build even stronger LLM judges that are friendly to local deployment.

7.3.2 Beyond Priority Language Averages

Table 12 shows a per-language breakdown of the systems included in Table 10, as well as the Lexicala systems, which were excluded from Table 10 because they do not cover all priority language pairs. While the top two systems generally achieve strong results across the board, a few systems outperform them on individual language pairs. For example, the Lexicala system achieves top-tier performance on eng-ekk and eng-sme, while PhraseT3-01 is competitive on eng-ind. A few systems are dragged down substantially by eng-sme, namely the Gemma 4 Reasoning and Qwen 3.6 No Reasoning baselines, which calls for more careful validation before deploying such systems on low-resource language pairs.

Table 13 expands Table 11 to all 21 language pairs, covering the subset of systems that support

Rank	Cluster	Submission	Underlying model	Track	MCC _{×100}	P(%)	R(%)	F1(%)
1	1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	33.40	50.00	53.15	49.84
2	2	Austriadactylus	Ensemble ^{E1}	T2	31.82	41.24	76.16	50.95
3	2	Pterodactyl	GPT-5.4	T2	31.70	41.55	72.94	51.13
4	2	Austriadraco	GPT-5.4	T3	31.67	42.54	69.56	50.99
5	3	Wukongopterus	GPT-5.4	T3	31.12	38.83	82.04	50.86
6	3	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	30.47	48.55	49.29	47.43
7	3	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	30.35	49.04	47.89	46.96
8	4	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	29.76	43.75	61.89	48.88
9	4	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	29.52	49.53	46.15	45.94
10	4	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	29.47	51.23	41.88	44.72
11	5	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	27.28	49.47	41.71	42.48
12	5	Cohere CAT+ ensemble	Command A+	T2	26.15	58.17	26.27	33.87
13	6	Cohere CAT+	Command A+	T1	23.16	45.39	38.43	39.17
14	6	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	21.41	51.79	23.61	30.03
15	6	First try [871721]	Qwen3.6-27B	T3	20.80	50.67	26.38	32.46
16	7	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	18.81	47.91	23.04	28.48
17	7	FACET	GPT-4.1 Mini (2025-04-14)	T1	18.81	47.88	23.06	28.50
18	7	gemba-poly	Llama 3.3 70B	T2	18.72	41.33	38.15	37.76
19	8	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	16.58	44.00	23.23	27.91
19	8	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	16.58	44.00	23.23	27.91
21	8	FACET	GPT-4.1 Mini (2025-04-14)	T2	16.56	43.94	23.25	27.92
21	8	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	16.56	43.94	23.25	27.92
23	8	First try [874374]	Babel-83B-Chat	T3	15.30	31.17	78.49	42.78
24	9	xCOMET XXL	xCOMET-XXL	T1-B	13.59	38.72	36.23	33.28
25	10	xCOMET XL	xCOMET-XL	T1-B	4.90	34.79	13.52	16.69

Table 13: WMT26 Task 3 ranking over all 21 language pairs, restricted to systems that cover every pair and macro averaged over pairs. Columns, tracks and row shading are as in Table 11. 13 further systems did not cover all 21 language pairs and are omitted; their results appear in the per-language tables. Cluster is defined as in Table 10, with the same test applied over all 21 pairs. Ensemble components named in the model cards are ^{E1} Austriadactylus: GPT-5.4; GPT-5.6 Terra.

them. While the most interesting trends are already well represented by the priority language pairs, the most notable change is that the Gemini 3.6 Flash baseline now overtakes the previously top-ranking Wukongopterus and Austriadraco systems and claims first place, with its advantage confirmed by statistical significance testing. The FACET systems also overtake First try [874374] in the ranking, potentially driven by the wider multilingual support of GPT-4.1 Mini (2025-04-14) compared with Babel-83B-Chat.

PTT-MULTITAN-v2 does not appear in any of the tables above, as it does not cover any of the priority language pairs. It is also the only non-LLM submission to this task. Detailed per-language results in the Appendix show that it performs largely on par with xCOMET-XL and xCOMET-XXL on the language pairs it covers.

7.3.3 Conversion Threshold from Task 2 Submissions

As mentioned above, one important variable in Task 2 to 3 conversion is the score threshold above which a segment is considered “error-free”. In

all previous analyses, we stick with the thresholds provided by the participants. Here, we investigate the optimality and sensitivity of these thresholds. Rather than adopting numerical optimization methods, we optimize thresholds with a deterministic sweep over all distinct submitted score boundaries, applying an inclusive conversion rule ($\geq$) without splitting tied scores. For each setup, we select the threshold that maximizes MCC, choosing a central cutpoint within tied optimal plateaus and favoring the plateau nearest the submitted threshold. Because selection and evaluation use the same assessments, these results should be treated as post-hoc hindsight rather than as independently validated improvements.

One choice that has to be made for such threshold optimization is whether individual thresholds are needed for each language pair, or whether a unified threshold for all language pairs (which is what we asked task participants to provide) is largely sufficient. To answer this question, we experiment with two versions of the optimization objective: in the first, we run the optimization described above

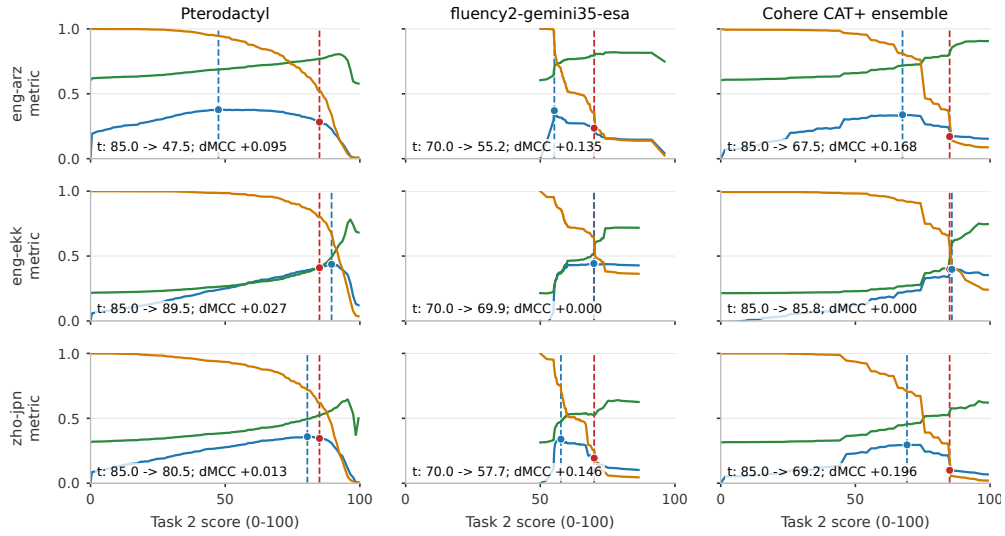
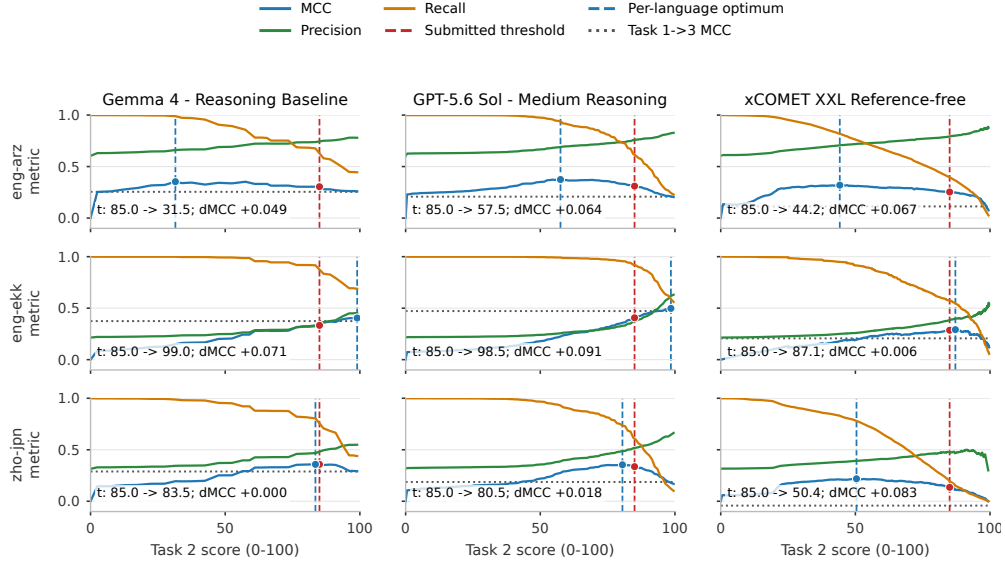


Figure 4: Task 2 to 3 conversion threshold sensitivity for selected organizer baselines (a) and participant submissions (b). Rows show eng_Latn-arz_Arab, eng_Latn-ekk_Latn, and zho_Hans-jpn_Jpan; columns span lower-precision/higher-recall to higher-precision/lower-recall submitted operating points. Curves give MCC (blue), precision (green), and recall (orange). Curves omit the extreme high-threshold tail and its final endpoint. Vertical red and blue dashed lines mark submitted and per-language optimal thresholds; annotations give threshold changes (t) and corresponding MCC changes ($dMCC$). Dotted gray lines in (a) give matching Task 1 to 3 MCC. Scores use the submitted 0–100 scale, error-free is positive, and predictions use score $\geq$ threshold.

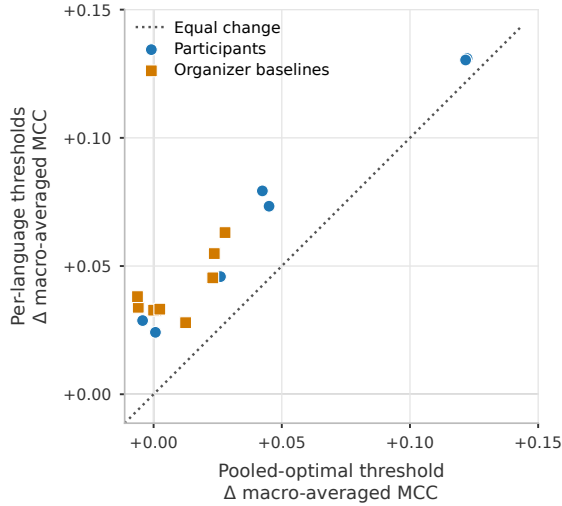


Figure 5: MCC impact of optimizing Task-2 score thresholds globally across all language-pairs versus separately for each language-pair, for the 21 systems covering all eight priority language-pairs. The horizontal axis shows micro-averaged MCC change from one global optimized threshold; the vertical axis shows macro-averaged MCC change from independently selected per-language thresholds. Circles are participant submissions and squares are organizer baselines; the dotted diagonal marks equal numerical change.

separately for each language pair; in the second, we run a single unified optimization on data points pooled across all language pairs, effectively optimizing micro-averaged MCC.¹⁴

Figure 5 compares per-language optimization with unified optimization. First, in the vast majority of cases, optimizing the threshold yields gains in macro-averaged MCC. In some cases (FACET and doc-FACET), the MCC gain is as large as 0.12. The few cases where optimization results in a drop are caused by the slight mismatch between micro- and macro-averaging of MCC during optimization and evaluation. Second, there is a gap of approximately 0.03 MCC between systems using a jointly optimized threshold and those using per-language optimized thresholds. The largest gap is 0.04 (from GPT-5.6 Terra). Multiple factors contribute to this gap, one of which is how well the scoring scales are calibrated across languages.

Figure 4 shows a selected set of organizer baselines and participant submissions on a few language pairs. The selection is made to give a good repre-

sentation of the different precision and recall profiles found among the converted systems. When measured by MCC, the optimal conversion threshold varies widely across language pairs and scoring systems. It is therefore not surprising that task participants often make substantially sub-optimal choices of conversion threshold for their own systems. This underscores the need to optimize scoring thresholds on a held-out set when deploying a scoring-based automatic evaluator for error-free translation detection, whenever a held-out set of human judgments is available.

Running this analysis on the baselines also allows us to compare Task 1 to 3 and Task 2 to 3 conversions under a setup with a controlled underlying system and prompt. Figure 4a shows that for all selected language $\times$ system combinations, there is at least one threshold that allows Task 2 to 3 conversion to outperform its Task 1 to 3 counterpart. In fact, when this analysis is expanded to all 9 baseline systems on the priority language pairs, only two combinations favor Task 1 to 3 conversion across the board. The size of the gap between the two conversion paths, however, varies considerably by language pair. So does the range of thresholds over which Task 2 to 3 conversion comes out ahead. It is also worth noting that for both xCOMET-XL and xCOMET-XXL, Task 1 to 3 conversion can hardly compete with its Task 2 to 3 counterpart on many language pairs.

8 Challenge Sets

For the fourth year, the shared task included a dedicated challenge set subtask. Inspired by the *Build it or break it: The Language Edition* shared task (Ettinger et al., 2017), the subtask complements evaluation on generic test data with targeted test sets designed to probe specific capabilities and failure modes of automated translation evaluation systems. Rather than aiming to represent naturally occurring translation data broadly, challenge sets focus on particular phenomena or difficult cases that may not be adequately captured by standard evaluation. The subtask is organized in a decentralized manner: the challenge sets are contributed by independent research teams, who define the phenomena to be tested and are subsequently responsible for analyzing and reporting the behavior of participating systems on their data.

¹⁴In principle, one could also optimize macro-averaged MCC directly, but this requires more complicated tracking of how much each individual data point affects the overall metric, which we did not implement.

Challenge Set	LPs	Phenomena	Items
AGGREE	6	Metric disagreement	350
FECE	3	Error correction evaluation	3000
GAMBIT+	7	Gender bias	9156
LIGATUR	2	Unicode robustness	171
QEBreak	10	QE scoring biases	2886

Table 14: Overview of the participation in the challenge sets subtask.

8.1 Subtask Structure

The challenge set subtask is organized into three stages: the *Breaking Round*, the *Scoring Round*, and the *Analysis Round*:

1. During the *Breaking Round*, participating teams (*Breakers*) submit challenge sets constructed to probe specific weaknesses or difficult phenomena in automated translation evaluation. A challenge set may address one or more of the three main subtasks: segment-level error detection and span annotation, segment-level quality score prediction, and detection of error-free segments. Depending on the target subtask, an example may contain one or more translation hypotheses and may include zero, one, or multiple reference translations.
2. In the *Scoring Round*, the challenge sets are added to the evaluation data of the relevant subtask(s). Participating evaluation systems (*Builders*) then produce the corresponding scores or predictions, while the organizers also obtain outputs from the baseline systems.
3. In the *Analysis Round*, the resulting system outputs are returned to the teams that submitted each challenge set. The *Breakers* then evaluate how the participating systems behave on the targeted phenomena and report their findings.

8.2 Challenge Set Descriptions and Individual Findings

This year, there were five submissions, covering a wide range of phenomena and language pairs. Table 14 provides an overview of the submitted challenge sets. A short description of every submission follows:

AGGREE (Griesbeck and Geierhos, 2026) is an aggregation-grounded challenge set for the diagnostic evaluation of segment-level MT metrics. It is motivated by the observation that high agreement between metrics at the system level can conceal substantial disagreement on individual segments.

The challenge set contains 350 naturally occurring system outputs across six language pairs drawn from WMT23 and WMT24, selected from cases where automatic metrics disagree locally or diverge from available human judgments. The examples were mined using diagnostics for inter-metric disagreement and metric-human mismatch and were then filtered using strict automated quality-control criteria. AGGREE is intended to characterize local metric reliability and disagreement patterns rather than to quantify performance degradation relative to standard test sets.

FECE (Ding and Duan, 2026) is a challenge set designed to evaluate the quality of fine-grained translation error correction. Existing MT evaluation metrics operate primarily at the segment level and may therefore fail to reflect whether a specific error has actually been corrected. This can lead to cases where edits outside a human-identified error span improve the overall metric score, while valid corrections to the annotated error itself receive little or even negative credit. FECE is constructed to expose this limitation by focusing on alternative corrections of fine-grained translation errors, allowing the authors to test whether metrics accurately reflect the effect of the correction itself.

FECE demonstrates that whole-segment metrics often fail to reward genuine error correction: across 29 WMT26 systems, average preference for the error-correcting candidate was just 51.2% for English–German, and metrics actually favored the merely-polished, error-preserving candidate for English–Japanese and English–Chinese. This shows that unrelated polishing can outweigh credit for the correction itself. Drawing from these observations, the authors proposed the COMET-Fine metric, which isolates only error-related edits and consistently favored the error-correcting candidate, reaching only 46–53% agreement against the WMT26 metrics.

GAMBIT+ (Menis Mastromichalakis et al., 2026) was introduced in WMT 2025 (Filandrianos et al., 2025) as a controlled challenge set for investigating gender bias in automatic translation evaluation. It is based on source segments containing gender-ambiguous occupational terms, paired with carefully matched masculine and feminine translations. Since the source does not specify the gender of the person associated with the occupation, both realizations are equally valid, and an unbiased evaluator should assign comparable scores to the two variants. For the 2026 edition, the authors

use an occupation-balanced subset of GAMBIT+ covering seven English-source language pairs with Arabic, Czech, German, Greek, Icelandic, Russian, and Ukrainian as target languages, extending the original resource with German. The subset contains 1,308 masculine/feminine translation pairs per target language, with three examples for each of the 436 ISCO-08 occupational groups¹⁵. This setup enables analysis of the direction, magnitude, and frequency of gender-related score differences across evaluators, languages, and occupations.

The results of GAMBIT+ show that gender bias remains present in MT evaluation, but that capturing its extent requires looking beyond a single aggregate measure to complementary dimensions of evaluator behavior.

LIGATUR (Mishra et al., 2026) is a 171-item contrastive challenge set that probes the sensitivity of MT evaluation metrics to Unicode format and control characters in English-German and English-Hindi. The set distinguishes between perturbations that should be effectively invisible to human readers and perturbations that alter the rendered text or its legibility. In particular, the authors examine Unicode characters whose normalization causes clean and corrupted strings to collapse to identical token sequences under commonly used metric encoders, including XLM-R and mT5, making the corruption inaccessible to the metric before scoring. Other perturbations instead fragment tokenization and can be detected more reliably. The English-Hindi portion additionally examines control characters affecting Devanagari conjunct formation, allowing the challenge set to distinguish between benign Unicode variation and character-level corruption with visible consequences.

A blind re-evaluation with held-out labels also overturns the authors’ pilot claim that judges ignore Devanagari conjunct breaks. As a result, the authors create the metric AEGIS (submitted in Task 2), which acts on the diagnosis, combining a reference-based ensemble with deterministic hygiene penalties for exactly what the encoders cannot see, and scores all 26,885 test items.

QEBreak (Silchenko, 2026) is a contrastive challenge set designed to audit systematic failure modes in machine translation quality metrics, with a particular focus on omission and addition asymmetry. It contains 2,886 segments across 10 language

pairs and comprises three controlled components: matched base, omission, addition triples that test whether metrics penalize silently omitted source content differently from added content; a fluent, fact-free verbosity ladder that probes sensitivity to length inflation; and minimal pairs containing numeric contradictions as a positive control.

Scores from 31 systems confirm the deficit and refute the QEBreak mechanism. Reference-free QE metrics under-penalize omission (a current QE baseline prefers the omission over its own base in 19.5% of matched base-versus-omission contests), and the edit-direction by reference-availability interaction is significant. But reference-based neural metrics are asymmetric in the same direction outside the pre-registered equivalence margin, so omission under-penalization belongs to learned metrics as a class; reference access softens it but does not remove it. Every returned metric penalizes pure verbosity, and two systems score corrupted numerals above correct ones.

8.3 Aggregated challenge set results

An aggregated view of the results provided by the challenge sets can be seen in Table 15. Here, the heterogeneous scores calculated by every challenge set have been normalized to form a ranking for each of its supported language pairs. This way, a direct comparison of how each system performed on each type of challenge is possible. Additionally, we calculate an aggregate rank, which is the result of the average of the individual challenge set ranks per language pair, weighted by the number of their respective test items. The ranks from Task 2 (Section 6.2) are also presented aligned with the aggregated ranks. It should be noted that the aggregated ranking is heavily skewed to the devised challenges and their language directions and does not represent the overall performance of the system, which is what we want to focus on through this comparison.

Performance is generally coherent across the language-pair columns belonging to the same challenge set. Correlation across challenge sets is weak to absent. Systems that dominate QEBreak are unremarkable on GAMBIT+ and LIGATUR, and the systems that struggle most on GAMBIT+ are not uniformly the worst elsewhere.

Strikingly, the challenge-set average ranking is largely inverted relative to the Task 2 ranking. Baseline lexical/embedding metrics (YiSi-1, BERTScore, spBLEU, BLEU, AEGIS,

¹⁵<https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/>

System	Rank	Task2	FECE			GAMBIT+							LIGATUR		QEBreak									
			eng-deu	eng-jpn	eng-zho	eng-ara	eng-ces	eng-deu	eng-ell	eng-isl	eng-rus	eng-ukr	eng-deu	eng-hin	ces-deu	ces-ara	eng-ara	eng-ces	eng-deu	eng-jpn	eng-kor	eng-rus	eng-zho	zho-jpn
items			1000	1000	1000	1308	1308	1308	1308	1308	1308	1308	89	70	108	252	88	251	607	31	33	231	138	10
YiSi-1	2.8	35	2	3	5	—	—	—	—	—	—	—	8	5	1	1	1	1	1	11	1	3	1	1
BERTScore	3.7	29	3	1	2	—	—	—	—	—	—	—	26	19	1	7	1	4	6	11	1	6	1	1
spBLEU	3.9	22	1	2	3	—	—	—	—	—	—	—	8	5	1	9	1	4	9	14	1	8	5	1
BLEU	4.4	24	4	4	1	—	—	—	—	—	—	—	4	7	1	14	1	2	6	21	1	9	7	1
AEGIS	4.7	23	10	15	10	1	3	1	1	4	3	3	1	1	9	5	7	7	1	7	14	9	9	14
COMETKiwi22*	5.5	39	11	9	15	2	2	2	2	3	1	1	18	7	15	9	25	16	11	1	15	16	15	1
chrF	5.8	17	7	5	4	—	—	—	—	—	—	—	4	7	1	8	1	2	11	11	1	3	7	1
bytepop_pro9	7.3	4	5	10	6	7	5	7	11	12	6	11	2	1	1	2	9	7	1	1	1	1	3	1
bytepop_pro10	7.7	3	5	10	7	9	5	7	11	13	6	11	2	1	1	2	7	7	1	1	1	1	3	1
COMET22	8.3	19	8	14	9	—	—	—	—	—	—	—	8	4	1	4	1	4	6	9	1	5	9	1
Vertical_1†	8.3	25	16	16	16	5	4	3	4	2	2	5	19	15	22	17	13	18	20	14	21	20	22	23
Vertical_2†	8.7	21	13	12	14	4	7	11	5	5	5	4	14	13	15	18	13	14	18	21	16	11	19	19
xCOMET XL†	9.7	31	12	7	8	10	9	5	—	9	8	7	7	—	17	22	23	20	19	1	11	17	19	1
gemba-poly*	11.1	26	28	28	28	3	1	4	3	1	4	2	28	18	27	26	24	24	27	24	18	24	25	18
xCOMET XL*†	11.3	40	14	13	13	10	9	5	—	10	8	7	23	—	24	21	21	23	22	7	11	22	17	1
Gemini-3.6-Flash*	12.8	1	18	20	17	15	8	12	—	6	10	6	22	—	17	15	17	17	16	16	18	21	23	19
xCOMET XXL†	14.1	27	9	6	11	17	18	9	—	18	20	20	8	—	11	6	9	10	16	1	10	6	13	1
doc-FACET†	14.5	36	22	25	25	6	11	15	8	7	11	9	16	10	22	27	21	25	23	24	25	23	24	23
FACET*	15.4	37	21	23	26	8	12	17	9	8	12	10	16	10	26	28	27	27	24	28	24	25	26	14
xCOMET XXL*†	15.7	38	15	8	12	16	18	9	—	18	20	20	12	—	19	9	26	22	21	1	1	18	14	1
Cohere CAT+ ensemble*	16.1	20	27	27	29	13	13	16	6	11	14	13	24	12	24	19	17	15	14	18	18	11	19	19
Gemma 4 - Reasoning	16.2	6	23	21	22	22	15	13	—	14	15	14	6	—	13	13	15	11	5	18	16	11	9	28
Gemma 4 - Reasoning*	16.3	10	25	19	19	21	16	14	—	14	16	14	13	—	12	12	17	12	9	9	11	11	9	19
MQM-LLM†	16.7	28	19	18	21	12	14	18	7	16	13	16	21	16	30	29	30	29	26	27	27	28	29	29
MQM-LLM Conf. Aware*†	17.5	30	17	17	18	14	17	19	10	17	17	17	20	17	29	25	29	28	25	24	27	26	28	29
cuni-v14-regression*	19.2	16	—	—	20	—	22	—	—	—	—	—	—	—	—	16	—	13	—	—	—	—	5	—
Qwen 3.6 - Reasoning*	20.0	13	24	22	24	19	20	21	—	20	18	18	25	—	20	20	17	19	14	18	22	18	18	23
Qwen 3.6 - Reasoning	20.1	11	26	24	23	18	21	20	—	20	19	18	14	—	20	23	15	21	13	21	22	15	15	23
Lexicala-QE-Ens (Majority)	21.3	—	—	—	—	23	—	—	—	—	—	—	—	—	10	—	11	—	—	—	—	—	—	14
fluency2-gemini35-esa*	22.1	33	20	25	27	20	23	22	13	22	22	22	26	14	28	24	28	26	28	16	26	27	26	27
Lexicala-QE-Ens (STAPLE)	22.5	—	—	—	—	24	—	—	—	—	—	—	—	—	14	—	12	—	—	—	—	—	—	14

Table 15: Indicative aggregated challenge set ranking calculated by the average of the individual challenge set ranks per language pair, weighted by the number of items. Comparison with the Task 2 ranks (Table 8) also shown. * marks reference-free submissions. † marks fine-tuned systems.

chrF) top the challenge-set average rank but sit mid-to-bottom on Task 2, while LLM judges (Gemini-3.6-Flash*, Gemma 4, Qwen 3.6) lead Task 2 but rank near the middle of the challenge-set aggregate. This reflects how each challenge set targets a failure mode tied to a specific system design: reference-free (*) judges under-penalize omission on QEBreak (lacking a reference to detect missing content), fine-tuned (†) neural metrics built on subword encoders (xCOMET, Vertical_1/2) are blind to LIGATUR perturbations that collapse to identical token sequences, and fluency-sensitive metrics reward FECE’s polish confound rather than genuine error correction—all failure modes largely absent from simple lexical overlap metrics, whose narrowness, however, limits their Task 2 human-correlation.

8.4 Summary of challenge set findings

The results of the challenge sets reveal systematic and distinct weaknesses in all current MT evaluation approaches. Evaluators that agree strongly at the **system level** often diverge considerably on individual segments, underscoring that aggregate agreement can mask local unreliability. Segment-

level metrics largely fail to reward genuine **error correction**, often favoring superficially polished but error-preserving translations over those that fix the annotated error, but highlighting error-related edits substantially improves this behavior. **Gender bias** persists in translation evaluation, though its magnitude and direction vary between evaluators, languages, and occupations, indicating that it cannot be easily captured by a single aggregate measure. Evaluation metrics based on subword encoders are vulnerable to **Unicode perturbations** that collapse to identical token sequences before scoring, but are more sensitive to perturbations that fragment tokenization or visibly disrupt legibility. Reference-free and, to a lesser extent, reference-based neural metrics under-penalize **content omission** relative to addition, a bias that access to references attenuates but does not eliminate, while verbosity and numerical inaccuracy are more consistently penalized.

Taken together, the performance of the evaluators on the challenge sets was largely uncorrelated with their overall strength and ranking, indicating that the failure modes exposed by the challenge sets are largely attributable to the evaluation sys-

tem’s architectural choices, reference-free scoring, subword tokenization, and fluency sensitivity.

9 Conclusions

This paper presented the findings of the WMT26 Shared Task on Automated Translation Quality Evaluation Systems. By evaluating fine-grained error detection, quality-score prediction, and error-free classification on a common multilingual test set, this year’s task provides a joint view of capabilities that have historically been evaluated separately.

Across all three quality formulations, prompted LLM-as-a-judge systems now substantially outperform all conventional lexical and supervised neural metrics included in our evaluation. At the same time, the results do not identify a single universally best evaluator: rankings remain dependent on model family, language direction, reference availability and provenance, and in some cases the use of reasoning. Open-weight models are competitive with proprietary frontier systems and strong reference-free evaluators perform particularly well. The new cross-task conversions further show that systems developed for fine-grained error detection or continuous quality scoring can often be converted effectively into error-free classifiers, with specialized Task 3 systems providing surprisingly little consistent advantage.

Several important questions remain open. Evaluation quality is still considerably less stable for lower-resource languages, and future work should investigate how to improve multilingual generalization without language-specific calibration. Furthermore, preliminary cross-domain analyses reveal an interesting pattern: although auto-evaluator rankings remain quite stable, evaluator performance varies considerably across domains, with higher scores in speech and software and lower scores in news, social, fact-checking, and educational domains. Future work should investigate these differences more thoroughly. The mixed effects of native, post-edited, and pseudo-references motivate further study of when references help and how evaluators should condition on reference provenance and estimated quality. The benefits of reasoning also appear to be strongly model-dependent, suggesting that evaluation accuracy should be considered jointly with inference cost rather than assuming that additional reasoning is uniformly beneficial. For error-free detection, robust threshold calibration across languages, domains, and different class dis-

tributions remains an important practical problem. More broadly, the strong performance of cross-task conversions raises the possibility of developing unified evaluators whose error spans, holistic scores, and accept/reject decisions are mutually consistent. Finally, the submitted challenge sets highlight multiple phenomena that remain poorly captured by aggregate benchmark performance and provide useful targets for future editions of the shared task.

Ethical Considerations

The data for this shared task was generated, screened, and human-annotated by the WMT26 General Machine Translation Shared Task. We acknowledge inheriting any ethical limitations and concerns raised by that task.

Generative-AI tooling was utilized by the task organizers in multiple aspects of the work reported here. Usage varied across organizers, with some utilizing AI more broadly while others did not use AI at all. Tables, figures and some analyses were generated by scripts that were largely automated using AI, but code was reviewed and results were cross-checked by the organizers. The content of this findings paper was mostly written natively by the committee members, with overview sections (abstract, introduction, conclusions) partly or mostly suggested by AI, and with AI assisting in reviewing and suggesting edits to the overall manuscript.

Acknowledgments

The execution of this shared task would not be possible without tight collaboration with the organizers of the WMT26 General MT Shared Task, whom we thank for their collaboration. We thank Google and OpenAI for providing the compute resources necessary for running the complete evaluation with several baseline auto-evaluators based on their proprietary models; and Google and the National Research Council of Canada for providing compute resources for generating the pseudo-references. We also thank Brigham Young University’s Office of Research Computing for the compute resources that made running the open-weight LLM baselines in this work possible.

Eleftherios Avramidis was partially funded by the German Federal Ministry of Research, Technology and Space (BMFTR) via the project RoGSiLT (grant number 16IW26001).

References

- Marina Iuliana Aur, Stefano Perrella, and Roberto Navigli. 2026. [Exploring the impact of document context on LLM-based machine translation evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2026*, Budapest, Hungary. Association for Computational Linguistics.
- Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. [Findings of the WMT 2023 shared task on quality estimation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). *Preprint*, arXiv:1911.02116.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2021a. [A statistical analysis of summarization evaluation metrics using resampling methods](#). *Transactions of the Association for Computational Linguistics*, 9:1132–1146.
- Daniel Deutsch, Rotem Dror, and Dan Roth. 2021b. [A statistical analysis of summarization evaluation metrics using resampling methods](#). *Transactions of the Association for Computational Linguistics*, 9:1132–1146.
- Daniel Deutsch, George Foster, and Markus Freitag. 2023. [Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929. Association for Computational Linguistics.
- Colten DiIanni and Daniel Deutsch. 2025. [Don’t sweat the small stuff: Segment-level meta-evaluation based on pairwise difference correlation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25062–25070, Suzhou, China. Association for Computational Linguistics.
- Jie Ding and Xiangyu Duan. 2026. FECE: A challenge set for evaluating fine-grained translation error correction. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Alan Eckhardt, Ondrej Glembek, and Craig Stewart. 2026. Phrase at WMT26 automated MT evaluation task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. [Towards linguistically generalizable NLP systems: A workshop and shared task](#). In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10. Association for Computational Linguistics.
- George Filandrianos, Orfeas Menis Mastromichalakis, Wafaa Mohammed, Giuseppe Attanasio, and Chrysoula Zerva. 2025. [GAMBIT+: A challenge set for evaluating gender bias in machine translation quality estimation metrics](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 314–326, Suzhou, China. Association for Computational Linguistics.
- Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. [MLQE-PE: A multilingual quality estimation and post-editing dataset](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974. European Language Resources Association.
- 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. Association for Computational Linguistics.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68. Association for Computational Linguistics.
- Evelyn Y. Garland and Carola F. Berger. 2026. Pterodactyl: Pushing the limits of efficient automated translation evaluation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Google DeepMind. 2026. Gemini 3.6 Flash model card. <https://deepmind.google/models/model-cards/gemini-3-6-flash/>.
- Lena Griesbeck and Michaela Geierhos. 2026. AG-GREE: An aggregation-grounded challenge set for segment-level MT metric evaluation. In *Proceedings of the Eleventh Conference on Machine Translation*,

- Budapest, Hungary. 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.
- Miroslav Hrabal and Ondřej Bojar. 2026. CUNI at the WMT26 automated MT evaluation shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Jisoo Jang, Haewon Cho, and Seungtaek Choi. 2026. HUFSDILAB at WMT 2026 automated translation quality evaluation task 1: Language-pair prompt routing for error span annotation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Diptesh Kanojia, Chi-kiu Lo, Archchana Sindhuja, Samuel Larkin, Greg Hanneman, and Alon Lavie. 2026. In the blind: Building pseudo-references for MT evaluation. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Ahrii Kim, Chanjun Park, and Seong-heum Kim. 2026. FACET at WMT 2026 automated translation quality evaluation task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026a. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Tom Kocmi, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Shaun Rafael Cassini, Ona de Gibert, Aidan Gomez, Nithya Govindarajan, Shun Kiyono, Olivia Lasche, Lawrence Rogers, Kelly Marchisio, Nikita Moghe, Yash More, Camila Moran-Hidalgo, Yiyang Nan, Michael Sachs, Trisha Starostina, Daan van Stigt, and 4 others. 2026b. North small translate: Advanced cost-effective translation (cohere CAT+). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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. 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. European Association for Machine Translation.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- 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.
- Soyoung Lee. 2026. PragmaSpan at WMT26: Taxonomy-guided few-shot error span detection and an empirical analysis of MPP. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Senyu Li, Jiayi Wang, Felermino DMA Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenertorp, and David Ifeoluwa Adelani. 2025. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced african languages? *arXiv preprint arXiv:2506.04557*.
- Chi-kiu Lo. 2019. [YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513. Association for Computational Linguistics.
- Boxuan Lyu, Haiyue Song, Hidetaka Kamigaito, Chenchen Ding, Hideki Tanaka, Masao Utiyama, Kōtarō Funakoshi, and Manabu Okumura. 2026. [Minimum bayes risk decoding for error span detection in reference-free automatic machine translation evaluation](#). *Preprint*, arXiv:2512.07540.
- Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

- Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Wafaa Mohammed, Giuseppe Attanasio, and Chrysoula Zerva. 2026. Benchmarking gender bias in machine translation evaluation metrics across occupations. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Animesh Mishra, Krishang Sharma, and Sonia Khetarpaul. 2026. Translation metrics cannot judge what their tokeniser deletes: LIGATUR and AEGIS submissions to the WMT26 shared task on automated translation quality evaluation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Airana Mongush and Dmitrii Pukhov. 2026. Algebras FluencyScore at WMT26 automated translation quality evaluation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. [Mt-ranker: Reference-free machine translation evaluation by inter-system ranking](#). *Preprint*, arXiv:2401.17099.
- Mariam Nakhlé, Raheel Qader, Marco Dinarelli, and Hervé Blanchon. 2026. Vertical: Quality estimation of machine translation from université grenoble alpes. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- OpenAI. 2026. GPT-5.6 system card. <https://openai.com/index/gpt-5-6-system-card/>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Stefano Perrella, Eric Morales Agostinho, and Hugo Zaragoza. 2026. [Span-level machine translation meta-evaluation](#). *Preprint*, arXiv:2603.19921.
- Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024a. [Beyond correlation: Interpretable evaluation of machine translation metrics](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714. Association for Computational Linguistics.
- Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024b. [Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.
- Benjamin Pong. 2026. Post-hoc correction of machine translation error span predictions. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.
- Lorenzo Proietti, Roman Grundkiewicz, and Matt Post. 2026. [PEAR: Pairwise evaluation for automatic relative scoring in machine translation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 42189–42207, San Diego, California, United States. Association for Computational Linguistics.
- Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025a. [Has machine translation evaluation achieved human parity? the human reference and the limits of progress](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–813. Association for Computational Linguistics.
- Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025b. [Estimating machine translation difficulty](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.
- Qwen Team. 2026. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585. Association for 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. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. Association for Computational Linguistics.
- Richard Servajean, Mohammad Golam Sohrab, Lucie Kunitomo-Jacquin, and Matiss Rikters. 2026. Leveraging verbalized confidence in LLM-as-a-judge for automated translation quality evaluation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Assaf Siani. 2026. Reference-free consensus evaluation of machine translation: Ensembling LLM error-span judgments with neutral cross-verification and STAPLE reliability fusion. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Maksim Silchenko. 2026. QEBreak at WMT26: A pre-registered audit of omission and addition asymmetry in machine translation quality metrics. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Archchana Sindhuja, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. [When LLMs struggle: Reference-less translation evaluation for low-resource languages](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Gemma Team. 2026. [Gemma 4 technical report](#). Preprint, arXiv:2607.02770.
- Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. [Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234. Association for Computational Linguistics.
- Jia Xu. 2026. SpinPop: A fast spin metric for the WMT26 metrics shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. [Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE?](#) In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). Preprint, arXiv:1904.09675.
- Lichao Zhu, Maria Zimina-Poirot, and Nicolas Froeliger. 2026. Beyond rankings: Insights from ALTAE’s PTT-MULTITAN submission to the WMT 2026 shared task on automated translation quality evaluation systems. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Vilém Zouhar, Roman Grundkiewicz, Sara Rajae, Parker Riley, Martin Popel, Rachel Bawden, Philipp Koehn, Marine Carpuat, and Tom Kocmi. 2026a. [Contrastive esa: Human evaluation of multiple translations at once](#). Preprint, arXiv:2607.26640.
- Vilém Zouhar and Tom Kocmi. 2026. [Pearmut: Human evaluation of translation made trivial](#). Preprint, arXiv:2601.02933.
- Vilém Zouhar, Julia Kreutzer, Alon Lavie, Tom Kocmi, Matt Post, Ondřej Bojar, and Mrinmaya Sachan. 2026b. [Dynamically allocating evaluation effort for model ranking](#). Preprint, arXiv:2608.03437.
- Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. [COMET-poly: Machine translation metric grounded in other candidates](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 887–904, Suzhou, China. Association for Computational Linguistics.

A Participant System Descriptions

A.1 Task 1

For each submission, we asked participants to also hand over a model card that includes a system description. Here, we report such system descriptions verbatim.

Cohere CAT+ (Kocmi et al., 2026b) Cohere CAT+ machine translation specific model is build on top of the Cohere Command A+ architecture which is a mixture-of-experts with 24 active parameters out of 218 total parameters. The utilized multistep training protocol: coarse SFT focussing also on non-machine translation capabilities, specialized MT SFT, followed by DPO, RLVR and finished with small targeted DPO. We utilize its instruction following capabilities to assess error spans and quality estimation with a variant of GEMBA-MQM prompt.

cuni-cat-v4 (Hrabal and Bojar, 2026) A two-stage MQM-style error detector based on the open-weight Gemma 4 31B instruction model with 4-bit AWQ quantization. The system uses few-shot prompting and one-pass self-refinement, predicts optional MQM categories. It uses source and candidate translation text without reference translations.

cuni-v14 (Hrabal and Bojar, 2026) A two-stage MQM-style error detector based on the open-weight Gemma 4 31B instruction model with 4-bit AWQ quantization. The system uses few-shot prompting and one-pass self-refinement to identify major and minor target-side errors and source omissions. It uses source and candidate translation text without reference translations.

FACET (Kim et al., 2026) FACET evaluates each translation through two separate prompting passes. Fluency looks at the target alone and accuracy at the source and target together, so each judgment sees only the context its error type needs: fluency does not get confused with source fidelity, and meaning errors are checked against the source rather than guessed from the target. The two passes are merged per segment and scored on a shared four-point severity scale, with one fixed model across both passes. This submission works entirely at the segment level, matching the granularity at which the task scores.

doc-FACET (Kim et al., 2026) FACET evaluates each translation through three separate prompting passes. Fluency looks at the target sentence alone, accuracy at the source and target together, and consistency at the whole document rather than one segment. Each pass sees only the context its error type needs, which keeps the judgments clean: fluency does not get confused with source fidelity, and document-level consistency catches cross-segment errors like the same name or term translated differently in different places, which segment-level scoring misses. The three passes are merged per segment and scored on a shared four-point severity scale, with one fixed model across all passes. We also provide a submission that excludes the consistency pass, since the current evaluation scheme scores at the segment level and does not capture document-level quality.

fluencyscore-multispan-v2 (Mongush and Pukhov, 2026) fluencyscore-multispan-v2: reference-free LLM judge (Gemini 3.5 Flash) prompted to enumerate all distinct error spans

per hypothesis (0..N) with major/minor severity following MQM-style conventions, plus an explicit omission field. Phrases are mapped to character spans with a deterministic Unicode-normalizing locator (exact / normalized / bounded-fuzzy tiers) with occurrence disambiguation; overlapping and adjacent same-severity spans are merged; whole-segment spans are restricted to garbled outputs. Empty or unscored hypotheses fall back to our prior single-span system. Covers official 21 LPs + challenge sets on mteval-test26.

fluencyscore-multispan-v2.1 (Mongush and Pukhov, 2026) fluencyscore-multispan-v2.1: two-pass judge (base pass + second-opinion pass on empty hypotheses + exhaustive pass on five under-covered LPs), span union with dedup; reference-free LLM judge (Gemini 3.5 Flash) prompted to enumerate all distinct error spans per hypothesis (0..N) with major/minor severity following MQM-style conventions, plus an explicit omission field. Phrases are mapped to character spans with a deterministic Unicode-normalizing locator (exact / normalized / bounded-fuzzy tiers) with occurrence disambiguation; overlapping and adjacent same-severity spans are merged; whole-segment spans are restricted to garbled outputs. Empty or unscored hypotheses fall back to our prior single-span system. Covers official 21 LPs + challenge sets on mteval-test26.

gemma4-31b-v4ref-struct-best (Jang et al., 2026) Gemma-4-31B-it used off-the-shelf (no SFT/fine-tuning) as a translation error-span judge. Prompt engineering only: a reference-aware prompt with 8 few-shot demonstrations, JSON guided-decoding output, and a rule-based language-ID pre-filter for wrong-language segments. Calibration is applied per language pair rather than globally — severity, expected error density, span extent, and under/over-flagging guidance are tuned separately for each pair, using HUMEVAL25-VAL gold where that pair has gold coverage and structural cross-pair signals where it does not. The submission selects, for each of the 8 pairs, the variant that scored best on the shared-task leaderboard among our runs.

Lexicala-QE-Ensemble (Majority-Vote) (Siani, 2026) A reference-free, LLM-ensemble quality-estimation system. Three independent LLM judges (Gemini 2.5 Flash, DeepSeek V4 Flash, Qwen Plus) annotate each translation under an MQM/ESA-style protocol; a neutral cross-

verification stage re-judges every proposed span, and a dedicated source-coverage stage detects omissions. The verified per-character judge masks are fused by CHARACTER-LEVEL MAJORITY VOTE (a character is marked erroneous when at least two of the three judges cover it), giving Task-1 error spans with minor/major severity and an omission label. The Task-2 quality score is an inverse-variance fusion of a reliability-weighted holistic estimator and a span-derived MQM-penalty estimator (span confidence = vote share; judge reliability = agreement with the consensus), clamped to [0,100]. The Task-3 error-free label is an adequacy-gated conjunction (no error spans, no omission, and holistic quality at or above threshold). No reference translations are used at any stage.

Lexicala-QE-Ensemble (STAPLE) (Siani, 2026) A reference-free, LLM-ensemble quality-estimation system. Three independent LLM judges (Gemini 2.5 Flash, DeepSeek V4 Flash, Qwen Plus) annotate each translation under an MQM/ESA-style protocol; a neutral cross-verification stage re-judges every proposed span, and a dedicated source-coverage stage detects omissions. Per-character judge masks are fused with STAPLE (Simultaneous Truth And Performance Level Estimation) — an EM algorithm that jointly estimates each judge’s sensitivity and specificity and the latent error truth, followed by HMM run-smoothing — yielding Task-1 error spans with minor/major severity and an omission label. The Task-2 quality score is an inverse-variance fusion of a reliability-weighted holistic estimator and a span-derived MQM-penalty estimator, clamped to [0,100]. The Task-3 error-free label is an adequacy-gated conjunction (no error spans, no omission, and holistic quality at or above threshold). No reference translations are used at any stage.

PhraseT1-01 (Eckhardt et al., 2026) Reference-free span-level (ESA) error detection for machine translation, using Gemini 3.5 Flash (Wmt26SpanTaskV2Directive prompt) as a single judge, no combination with other systems.

PhraseT1-02 (Eckhardt et al., 2026; Hrabal and Bojar, 2026) Ensemble span-level (ESA) error detection for machine translation, combining Gemini 3.5 Flash (closed-weight), Gemma 4 12B (open-weight, Phrase-served via local vLLM), and UFAL cat-v4 (Charles University’s two-stage MQM-style

error detector, based on open-weight Gemma 4 31B with 4-bit AWQ quantization, few-shot prompting and one-pass self-refinement). Combined via unweighted union voting across all three systems’ spans (a span counts if any system flagged it), with a Gemini+Gemma-only fallback for the small fraction of segments UFAL’s system did not cover.

PragmaSpan (Lee, 2026) PragmaSpan is a training-free, few-shot GPT-5.5 span-error detector wrapped in a reliability-oriented inference and submission harness. References are included as optional prompt context when provided, and provided ASR transcripts are used as text context for speech items; video and screenshots are not consumed. The harness converts structured API responses into complete, deterministic, format-valid WMT26 Task 1 artifacts through truncation-aware row retries, conservative JSON and offset recovery, deterministic exact-span deduplication, exact item/system completeness checks against the official input, and mandatory validation before packaging. These safeguards target artifact reliability rather than claiming to improve diagnostic accuracy. The detector uses a seven-axis diagnostic taxonomy covering semantic, linguistic, discourse, and interpersonal errors; an ablation found that its four pragmatic axes provide interpretive structure but change detection accuracy by at most ± 0.001 F on the evaluated domain mix. Model and prompt choices were evaluated against WMT25 ESA surrogate gold and WMT22 MQM en-de using the official severity-weighted MPP metric, and several prompt, span, and severity refinements were tested but excluded after failing to improve consistently across benchmarks.

Qwen-based Omission (Servajean et al., 2026) To address the Segment-Level Error Detection and Span Annotation, a reasoning-based LLM with few-shot learning that leverage structure output is applied. The Qwen3 32B model is adapted to generate the output that includes error span detection, classify each span according to its severity, and identify a possible omission for each segment in a completed translation.

Rule-based Omission (Servajean et al., 2026) We follow the same procedure for Qwen-based Omission except the omission is computed using a rule-based approach on the basis of overall severity of a completed translation.

QwenSpan [Citation missing] A reference-free multilingual translation error span detection system based on Qwen3-14B. The English-German system is trained with supervised fine-tuning followed by GRPO. The English-Chinese system is further trained on a mixture of English-German and English-Chinese data. The system takes the source sentence and machine translation as input and predicts minor and major error spans without using the provided reference translations.

XCOMOffset (Pong, 2026) This submission refines the state-of-the-art xcomet metric by rebalancing the class labels based on the label distribution of the training data. It modifies xcomet by adding an offset to the logits, hence adjusting the softmax subword probabilities for each error class. It then ranks the error spans based on the model’s entropy, selecting the top-k error spans to be re-classified or re-segmented by a reasoning LLM.

A.2 Task 2

AEGIS (Mishra et al., 2026) AEGIS is a reference-skeptical, corruption-shielded evaluation system. Motivated by WMT26’s shift to post-edited and synthetic references, it treats the reference as just another hypothesis: a reference-based COMET-22 score and a reference-free MetricX-24 (QE-mode) score are fused with a per-segment weight derived from the reference’s own estimated quality and declared type (human / postedit / pseudo / absent), falling back to pure QE where no reference exists. Before neural scoring, a deterministic Unicode-hygiene layer separates orthographic integrity from semantic quality: default-ignorable codepoints, whitespace-class substitutions, homograph cloaking, and Indic conjunct-breaking joiners are detected and converted into transparent point penalties (we verified XLM-R-based metrics are provably blind to several of these corruptions), and a degeneracy guard catches empty, runaway, and repetition-loop outputs. We submit predictions for every official language pair, every system, and all challenge-set items.

Austriadactylus, a variant of Pterodactyl (Garland and Berger, 2026) Austriadactylus, a variant of Pterodactyl, is a highly customizable, ultra-low-cost system optimized for binary classification in translation triage and data curation. It is based on an LLM-as-a-judge approach, with prompt engineering and model selection guided by ROC analysis. Unlike Pterodactyl, Austriadactylus uses GPT-

5.6 for ultra-low-resource languages (Ladin, Ligurian, and Northern Sami).

bytepop_pro10 (Xu, 2026) PRIMARY Ensemble + Gemini-3-Pro on all covered normal columns; Ladin=Gemini+sol; Lig/Sami=Gemini

bytepop_pro9 (Xu, 2026) Ensemble + Gemini-3-Pro on 8 losing + cs-de (full); Ladin=Gemini+sol; Lig/Sami=Gemini

Cohere CAT+ ensemble (Kocmi et al., 2026b) Cohere CAT+ machine translation specific model is build on top of the Cohere Command A+ architecture which is a mixture-of-experts with 24 active parameters out of 218 total parameters. The utilized multistep training protocol: coarse SFT focussing also on non-machine translation capabilities, specialized MT SFT, followed by DPO, RLVR and finished with small targeted DPO. We utilize its instruction following capabilities to assess error spans and quality estimation with a variant of GEMBA-MQM prompt. The final score is average of three completions from the same model and the same prompt: one greedy pass at temperature 0 plus two sampled passes at temperatures 0.3.

cuni-da-6d (Hrabal and Bojar, 2026) A direct segment-level DA/ESA system based on the open-weight Gemma 4 31B instruction model with 4-bit AWQ quantization. It consumes source text, candidate translation, language information, and the provided translation instruction, but does not use references. Its frozen DSPy prompt includes six labeled demonstrations selected using WMT25 training data before WMT26 test inference and covers all translations in the eight WMT26 high-priority language pairs.

cuni-v14-regression (Hrabal and Bojar, 2026) A frozen supervised regression model over aggregate error features produced by a two-stage MQM-style Gemma 4 31B detector (cuni-v14 task 1 submission). The constrained weighted-error regression was fit and selected only on WMT25 training data using underlying-document grouped validation, excludes prediction failures from fitting, and uses a frozen train-mean fallback. It does not use reference translations. This submission covers the eight WMT26 high-priority language pairs plus English-Czech, Czech-Ukrainian, and all released English-Chinese variants.

doc-FACET (Kim et al., 2026) FACET evaluates each translation through three separate prompting

passes. Fluency looks at the target sentence alone, accuracy at the source and target together, and consistency at the whole document rather than one segment. Each pass sees only the context its error type needs, which keeps the judgments clean: fluency does not get confused with source fidelity, and document-level consistency catches cross-segment errors like the same name or term translated differently in different places, which segment-level scoring misses. The three passes are merged per segment and scored on a shared four-point severity scale, with one fixed model across all passes. We also provide a submission that excludes the consistency pass, since the current evaluation scheme scores at the segment level and does not capture document-level quality.

FACET (Kim et al., 2026) FACET evaluates each translation through two separate prompting passes. Fluency looks at the target alone and accuracy at the source and target together, so each judgment sees only the context its error type needs: fluency does not get confused with source fidelity, and meaning errors are checked against the source rather than guessed from the target. The two passes are merged per segment and scored on a shared four-point severity scale, with one fixed model across both passes. This submission works entirely at the segment level, matching the granularity at which the task scores.

fluency2-gemini35-esa (Mongush and Pukhov, 2026) fluency2-gemini35-esa: Gemini 3.5 Flash fluency2 judge with fluency201 ($\gamma=1.8$, cap 9) display transform; ESA via WMT25 linear calib on fluency201 + moment-match to WMT25 score marginal (avoids ESA=100 saturation). Official 21 LPs + challenge. Task3 threshold 85 inclusive. Unscored hyps (~4% official) filled with ESA=50.0 fallback.

gemba-poly (Züfle et al., 2025) GEMBA-poly-cand is a reference-free, LLM-based quality estimation metric built on Llama 3.3 70B. Each translation is scored alongside up to 5 additional candidate translations of the same source segment, selected by cosine similarity over all-MiniLM-L12-v2 sentence embeddings, giving the model contrastive context rather than scoring the translation in isolation. Candidate pools are drawn primarily from other participating systems’ outputs for the same source segment; for segments with too few available system outputs, additional synthetic

candidates were generated with TranslateGemma-4B to ensure a sufficient pool for similarity-based selection.

Lexicala-QE-Ensemble (Majority-Vote) (Siani, 2026) A reference-free, LLM-ensemble quality-estimation system. Three independent LLM judges (Gemini 2.5 Flash, DeepSeek V4 Flash, Qwen Plus) annotate each translation under an MQM/ESA-style protocol; a neutral cross-verification stage re-judges every proposed span, and a dedicated source-coverage stage detects omissions. The verified per-character judge masks are fused by CHARACTER-LEVEL MAJORITY VOTE (a character is marked erroneous when at least two of the three judges cover it), giving Task-1 error spans with minor/major severity and an omission label. The Task-2 quality score is an inverse-variance fusion of a reliability-weighted holistic estimator and a span-derived MQM-penalty estimator (span confidence = vote share; judge reliability = agreement with the consensus), clamped to [0,100]. The Task-3 error-free label is an adequacy-gated conjunction (no error spans, no omission, and holistic quality at or above threshold). No reference translations are used at any stage.

Lexicala-QE-Ensemble (STAPLE) (Siani, 2026) A reference-free, LLM-ensemble quality-estimation system. Three independent LLM judges (Gemini 2.5 Flash, DeepSeek V4 Flash, Qwen Plus) annotate each translation under an MQM/ESA-style protocol; a neutral cross-verification stage re-judges every proposed span, and a dedicated source-coverage stage detects omissions. Per-character judge masks are fused with STAPLE (Simultaneous Truth And Performance Level Estimation) — an EM algorithm that jointly estimates each judge’s sensitivity and specificity and the latent error truth, followed by HMM run-smoothing — yielding Task-1 error spans with minor/major severity and an omission label. The Task-2 quality score is an inverse-variance fusion of a reliability-weighted holistic estimator and a span-derived MQM-penalty estimator, clamped to [0,100]. The Task-3 error-free label is an adequacy-gated conjunction (no error spans, no omission, and holistic quality at or above threshold). No reference translations are used at any stage.

MQM-LLM (Servajean et al., 2026) Our approach uses an LLM-as-a-judge within the MQM

error typology. Error severities are used as inputs to train an MLP regressor to predict human annotations.

MQM-LLM Confidence Aware (Servajean et al., 2026) Our approach uses an LLM-as-a-judge within the MQM error typology, with the novel addition of asking the LLM to verbalize its confidence in each error type assessment. Error severities and their associated confidence ratings are used as inputs to train an MLP regressor to predict human annotations.

PhraseT2-01 (Eckhardt et al., 2026) Segment-level quality scoring (ESA-style 0-100 score) combining two LLM judges: Gemini 3.5 Flash (closed-weight) and Gemma 12B (open-weight, locally served via vLLM), both prompted with the same reference-free error-span prompt (Wmt26SpanTaskV2Directive) and their derived scores combined via a scalar RRWA-weighted mean using a metric-specific (T2 Pearson) static reliability prior per judge.

PhraseT2-02 (Eckhardt et al., 2026; Hrabal and Bojar, 2026) Segment-level quality scoring (ESA-style 0-100 score) combining Gemini 3.5 Flash (closed-weight), Gemma 4 12B (open-weight, Phrase-served via local vLLM), and UFAL cat-v4 (Charles University’s two-stage MQM-style error detector, based on open-weight Gemma 4 31B with 4-bit AWQ quantization, few-shot prompting and one-pass self-refinement), via a scalar mean of each system’s own score weighted by a per-system reliability prior (each system’s standalone Pearson correlation on held-out WMT25 dev data), with a Gemini+Gemma-only fallback for segments UFAL did not cover.

Pterodactyl (Garland and Berger, 2026) Pterodactyl is a highly customizable, ultra-low-cost system optimized for binary classification in translation triage and data curation. It is based on an LLM-as-a-judge approach, with prompt engineering and model selection guided by ROC analysis.

Vertical_1 (Nakhlé et al., 2026) Vertical is an MT evaluation metric trained using an LLM as backbone (namely google/gemma-3-1b-it). We insert a special token into the input sequence, and the model uses it as the representation of the full sequence; hidden states across all layers are combined via a layer-wise attention mechanism before being fed to a two-layer linear network that pre-

dicts the final score on a 0 to 1 scale. We use the unified training approach, meaning we train the model on reference-based and reference-free data and leverage the decoder-only causal attention to predict the score under two variants: using the source and hypothesis as inputs (the special token is inserted after the hypothesis, so it only attends to the previous sequence) and using the source, hypothesis and reference as inputs (a second special token is inserted after the reference, so it attends to all previous tokens). The model is trained to minimize the weighted average of the MSE losses of the two variants against the ground truth score. At inference time, we use the appropriate variant depending on whether the reference is available, but any post-hoc combination of the two variant scores is possible when both exist. The model is trained on public WMT data from previous years. To boost performance on longer inputs, we propose a training curriculum that splits data into three stages: 1) sentence-level data, 2) paragraphs reconstructed from sentences, and 3) real paragraphs (WMT25 data).

Vertical_2 (Nakhlé et al., 2026) Vertical is an MT evaluation metric trained using an LLM as backbone (namely google/gemma-3-1b-it). We insert a special token into the input sequence, and the model uses it as the representation of the full sequence; hidden states across all layers are combined via a layer-wise attention mechanism before being fed to a two-layer linear network that predicts the final score on a 0 to 1 scale. We use the unified training approach, meaning we train the model on reference-based and reference-free data and leverage the decoder-only causal attention to predict the score under two variants: using the source and hypothesis as inputs (the special token is inserted after the hypothesis, so it only attends to the previous sequence) and using the source, hypothesis and reference as inputs (a second special token is inserted after the reference, so it attends to all previous tokens). The model is trained to minimize the weighted average of the MSE losses of the two variants against the ground truth score. At inference time, we use the appropriate variant depending on whether the reference is available, but any post-hoc combination of the two variant scores is possible when both exist. The model is trained on public WMT data from previous years. To boost performance on longer inputs, we propose a training curriculum that splits data into three stages: 1)

sentence-level data, 2) paragraphs reconstructed from sentences, and 3) real paragraphs (WMT25 data).

A.3 Task 3

Austriadraco and Wukongopterus (Garland and Berger, 2026) 1020 Group submitted two systems. Pterodactyl is a highly customizable, ultra-low-cost system optimized for binary classification in translation triage and data curation. It is based on an LLM-as-a-judge approach, with prompt engineering and model selection guided by ROC analysis. The first system is Austriadraco. Austriadraco is derived from Pterodactyl and specifically optimized for Task 3, employing a threshold-setting strategy based on one set of patterns observed in the WMT25 data. The second system is Wukongopterus. Wukongopterus is derived from Pterodactyl and specifically optimized for Task 3, employing a threshold-setting strategy based on one set of patterns observed in the WMT25 data.

First try (Servajean et al., 2026) AIST submitted two systems, both named First try (submissions 871721 and 874374), with the same description in their model cards. Their approach uses an LLM-as-a-judge within the MQM error typology, with the novel addition of asking the LLM to verbalize its confidence in each error type assessment.

PTT-MULTITAN-v2 (Zhu et al., 2026) ALTAE, Université Paris Cité submitted two systems, both named PTT-MULTITAN-v2 (submissions 875519 and 875520), with the same description in their model cards. We used a bottom-up approach that targets the weakness of the MCC metric. After fine-tuning a LightGBM classifier with WMT25’s data sets and applying data filtering strategies and CometKiwi filtering, we used multifactorial phraseology heuristics based on corpus linguistics and collocation measures: entropy, logDice, MI filtering, etc. We also used an LLM-as-a-judge strategy, further specializing the prompt by highlighting the disagreements between error-free segments and the classifier’s predictions. The system covers six language pairs: eng-bel, eng-deu, eng-rus, eng-ukr, eng-zho_CN and eng-zho_TW.

Lexicala (Siani, 2026) Lexicala submitted two systems. A reference-free, LLM-ensemble quality-estimation system. Three independent LLM judges (Gemini 2.5 Flash, DeepSeek V4 Flash, Qwen Plus) annotate each translation under

an MQM/ESA-style protocol; a neutral cross-verification stage re-judges every proposed span, and a dedicated source-coverage stage detects omissions. The first system is Lexicala-QE-Ensemble (Majority-Vote). The verified per-character judge masks are fused by character-level majority vote (a character is marked erroneous when at least two of the three judges cover it), giving Task-1 error spans with minor/major severity and an omission label. The Task-2 quality score is an inverse-variance fusion of a reliability-weighted holistic estimator and a span-derived MQM-penalty estimator (span confidence = vote share; judge reliability = agreement with the consensus), clamped to [0, 100]. The second system is Lexicala-QE-Ensemble (STAPLE). Per-character judge masks are fused with STAPLE (Simultaneous Truth And Performance Level Estimation) — an EM algorithm that jointly estimates each judge’s sensitivity and specificity and the latent error truth, followed by HMM run-smoothing — yielding Task-1 error spans with minor/major severity and an omission label. The Task-2 quality score is an inverse-variance fusion of a reliability-weighted holistic estimator and a span-derived MQM-penalty estimator, clamped to [0, 100]. Both descriptions close identically: The Task-3 error-free label is an adequacy-gated conjunction (no error spans, no omission, and holistic quality at or above threshold). No reference translations are used at any stage.

RAGthoven Error-Free [Citation missing] This submission comes from NaiveNeuron. A fixed reference-free pipeline for sentence-level error-free translation detection. It combines deterministic structural evidence with structured source-and-hypothesis judging, bounded independent verification, and deterministic aggregation.

PhraseT3-01 and PhraseT3-02 (Eckhardt et al., 2026; Hrabal and Bojar, 2026) Phrase submitted two systems. The first system is PhraseT3-01. Binary error-free segment classification combining two LLM judges: Gemini 3.5 Flash (closed-weight) and Gemma 12B (open-weight, locally served via vLLM), both prompted with the same reference-free error-span prompt (Wmt26SpanTaskV2Directive). A segment is classified error-free only if both judges independently agree (unweighted intersection voting on judge-detected spans). The second system is PhraseT3-02, submitted as Phrase + CUNI. Binary error-free segment classification combining Gemini

3.5 Flash (closed-weight), Gemma 4 12B (open-weight, Phrase-served via local vLLM), and UFAL cat-v4 (Charles University’s two-stage MQM-style error detector, based on open-weight Gemma 4 31B with 4-bit AWQ quantization, few-shot prompting and one-pass self-refinement), via a per-segment reciprocal-rank-weighted average (RRWA) of each system’s own error-free prediction, weighted by a per-system reliability prior (each system’s standalone MCC on held-out WMT25 dev data), with a Gemini+Gemma-only fallback for segments UFAL did not cover.

FACET (Kim et al., 2026) Soongsil University submitted two systems, both named FACET (submissions 869556 and 869557), with the same description in their model cards. FACET evaluates each translation through three separate prompting passes. Fluency looks at the target sentence alone, accuracy at the source and target together, and consistency at the whole document rather than one segment. Each pass sees only the context its error type needs, which keeps the judgments clean: fluency does not get confused with source fidelity, and document-level consistency catches cross-segment errors like the same name or term translated differently in different places, which segment-level scoring misses. The three passes are merged per segment and scored on a shared four-point severity scale, with one fixed model across all passes.

B Per-Language Pair Results

B.1 Task 1

Tables 16–36 report per-translation direction rankings. Each ranking includes the auto-evaluators that produced assessments for all translations in that translation direction.

B.2 Task 2

Tables 37–57 report Task 2 results for every official translation direction, including directions outside the priority set. Each ranking includes only direct Task 2 participant submissions and direct Task 2 baselines that produced scores for every translation with human scores in that direction. The eligible auto-evaluator set can therefore differ by direction. Task 1-to-Task 2 conversions are not included in these tables.

C Analysis of Task 2 Meta-Metrics

This appendix validates the grouping choices in the Task 2 meta-evaluation. All analyses use the

eight primary translation directions. They include only direct Task 2 participant submissions and direct Task 2 baselines with valid predictions for the relevant direction; automatically converted Task 1 submissions are excluded.

Screen-based versus source-based PDP.

Source-based PDP averages all retained human scores for an item–system pair and considers every pair of systems with a human score on that source item. Screen-based PDP instead retains only the pairs co-displayed in cESA screens. Table 58 reports the Spearman correlation between the corresponding auto-evaluator rankings for every primary direction. The PDP correlations are uniformly high ($\rho = 0.987$ – 0.999), showing that screen-based PDP closely recovers the source-based ranking while relying only on the direct contrastive human comparisons. We consequently select screen-based PDP as the official segment-level component, since it allows avoiding comparisons whose human scores were elicited in different contrastive contexts.

Screen-based versus source-based SPA.

The analogous SPA correlations in Table 58 are lower and more variable ($\rho = 0.775$ – 0.963). The difference follows from the evidence available to each MT system pair. Screen-based SPA can use only source items where the two systems were co-annotated, whereas source-based SPA uses every source item for which both systems have a retained human score. Figure 6 shows the resulting support gap: the median co-screen support ranges from 4 to 27 source items per pair, compared with 41 to 433 for source-based SPA. The gap persists for mean support and is especially acute in directions with sparse cESA pair overlap. To base the system-level meta-evaluation on substantially greater support, we therefore select source-based SPA as the official system-level component.

D Additional reference analyses for Task 2

Tables 59–62 provide the reference analyses discussed in Section 6.3.2.

Auto-evaluator	Ref-free	Ref-based	Δ Avg Corr
Gemma 4 [#]	0.800	0.814	+0.014
Qwen 3.6 [#]	0.786	0.800	+0.014
Gemma 4	0.794	0.808	+0.014
Qwen 3.6	0.742	0.764	+0.022
xCOMET XL	0.558	0.661	+0.103
xCOMET XXL	0.586	0.681	+0.095

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4	Gemma 4 31B IT	28.26	33.34	30.59
2	1	Gemma 4 - thinking	Gemma 4 31B IT	33.11	28.26	30.50
3	1	PhraseT1-01*	Gemini 3.5 Flash	37.34	25.35	30.20
4	1	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	30.21	30.04	30.12
5	2	fluencyscore-multispan-v2*	Gemini 3.5 Flash	41.28	23.03	29.57
6	2	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	40.75	23.17	29.55
7	2	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	33.01	26.56	29.43
8	3	Gemma 4*	Gemma 4 31B IT	29.56	27.70	28.60
9	3	GPT-5.6 Terra*	GPT-5.6 Terra	33.64	24.82	28.57
10	3	Cohere CAT+*	Command A+	29.72	27.36	28.49
11	3	Gemma 4 - thinking*	Gemma 4 31B IT	33.89	24.24	28.26
12	3	Gemini-3.6-Flash*	Gemini 3.6 Flash	40.49	21.35	27.96
13	3	GPT-5.6 Sol - xhigh thinking*	GPT-5.6 Sol	29.91	26.16	27.91
14	3	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	23.55	34.21	27.90
15	4	PhraseT1-02*	Ensemble ^{E1}	23.16	33.64	27.43
16	4	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	26.94	27.48	27.21
17	5	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E2}	23.30	30.70	26.49
18	5	cuni-cat-v4*	Gemma 4 31B	33.99	20.72	25.75
19	6	cuni-v14*	Gemma 4 31B	33.51	20.56	25.48
20	7	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	18.81	30.59	23.29
21	7	Qwen 3.6	Qwen3.6-35B-A3B	15.71	41.02	22.72
22	8	Qwen 3.6*	Qwen3.6-35B-A3B	15.45	35.70	21.56
23	9	doc-FACET*	GPT-4.1 Mini (2025-04-14)	18.79	20.75	19.72
24	9	FACET*	GPT-4.1 Mini (2025-04-14)	18.89	20.62	19.72
25	9	xCOMET XXL [†]	xCOMET-XXL	15.14	25.53	19.01
26	10	xCOMET XL [†]	xCOMET-XL	13.09	24.71	17.11
27	11	Qwen-based Omission ^{*†}	Qwen3-32B	13.44	12.21	12.80
28	11	Rule-based Omission ^{*†}	Qwen3-32B	13.43	12.18	12.77
29	11	xCOMET XL ^{*†}	xCOMET-XL	7.76	27.46	12.09
30	12	xCOMET XXL ^{*†}	xCOMET-XXL	6.60	21.41	10.09

Table 16: Task 1 ranking for the Czech→German translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus].

Table 59: Effect of reference use in Task 2. Avg Corr is the arithmetic mean of screen-based PDP and source-based SPA. # marks reasoning auto-evaluators. Light-grey rows are baselines.

Auto-evaluator	Main human	PseudoRef	Δ Avg Corr
bytepop_pro10	0.838	0.831	−0.006
bytepop_pro9	0.833	0.826	−0.007
Gemma 4 Reasoning	0.833	0.835	+0.003
Gemma 4 No Reasoning	0.828	0.828	+0.000
Qwen 3.6 Reasoning	0.807	0.804	−0.003
Qwen 3.6 No Reasoning	0.775	0.784	+0.009
Vertical_2	0.766	0.765	−0.002
chrF	0.765	0.751	−0.015
xCOMET XXL	0.754	0.751	−0.004
spBLEU	0.753	0.734	−0.019
COMET22	0.752	0.739	−0.014
xCOMET XL	0.744	0.733	−0.010
BLEU	0.741	0.723	−0.018
AEGIS	0.737	0.727	−0.010
Vertical_1	0.735	0.739	+0.004
BERTScore	0.682	0.679	−0.003
YiSi-1	0.626	0.619	−0.007

Table 60: Task 2 ranking under the main human and PseudoRef references. Each value is the equal-direction mean of screen-based PDP and source-based SPA over 4 matched priority directions. The source items, non-reference MT outputs, and human evidence are identical in both columns. Light-grey rows are organizer baselines. In this aggregate, the main human reference is from scratch for Czech to German, English to Estonian, and English to Icelandic, and post-edited for Chinese to Japanese.

Auto-evaluator	From scratch	Post-edited	Δ Avg Corr
Gemma 4 Reasoning	0.768	0.766	−0.002
bytepop_pro10	0.767	0.764	−0.002
bytepop_pro9	0.767	0.764	−0.002
Gemma 4 No Reasoning	0.760	0.756	−0.004
Qwen 3.6 Reasoning	0.757	0.741	−0.016
Qwen 3.6 No Reasoning	0.723	0.719	−0.004
chrF	0.700	0.701	+0.002
Vertical_2	0.695	0.694	−0.001
COMET22	0.687	0.686	−0.001
spBLEU	0.679	0.686	+0.007
BLEU	0.672	0.678	+0.006
AEGIS	0.661	0.667	+0.005
Vertical_1	0.660	0.662	+0.003
xCOMET XXL	0.652	0.647	−0.005
xCOMET XL	0.645	0.649	+0.003
BERTScore	0.634	0.639	+0.006
YiSi-1	0.592	0.597	+0.005

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	fluencyscore-multispan-v2 [*]	Gemini 3.5 Flash	32.05	26.82	29.20
2	1	fluencyscore-multispan-v2.1 [*]	Gemini 3.5 Flash	31.76	26.97	29.17
3	2	Gemini-3.6-Flash [*]	Gemini 3.6 Flash	30.44	26.52	28.34
4	3	Gemma 4 [*]	Gemma 4 31B IT	24.20	29.33	26.52
5	3	Gemma 4 - thinking [*]	Gemma 4 31B IT	26.64	26.00	26.32
6	4	cuni-cat-v4 [*]	Gemma 4 31B	25.72	22.11	23.78
7	4	cuni-v14 [*]	Gemma 4 31B	25.76	22.05	23.76
8	4	Qwen 3.6 - thinking [*]	Qwen3.6-35B-A3B	20.50	27.82	23.60
9	4	Cohere CAT+ [*]	Command A+	22.32	24.69	23.44
10	5	FACET [*]	GPT-4.1 Mini (2025-04-14)	15.48	26.38	19.51
11	5	doc-FACET [*]	GPT-4.1 Mini (2025-04-14)	15.45	26.41	19.49
12	6	Qwen 3.6 [*]	Qwen3.6-35B-A3B	12.32	37.79	18.58
13	7	xCOMET XL ^{*†}	xCOMET-XL	7.60	31.69	12.26
14	8	xCOMET XXL ^{*†}	xCOMET-XXL	6.39	26.25	10.28

Table 17: Task 1 ranking for the Czech→Ukrainian translation direction. ^{*} marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Table 61: Task 2 ranking under Human (from scratch) and Human (postediting) references. Each value is the equal-direction mean of screen-based PDP and source-based SPA over 2 matched priority directions. The source items, non-reference MT outputs, and human evidence are identical in both columns. Light-grey rows are organizer baselines.

Auto-evaluator	Ref-free	Ref-based	Δ score
Gemma 4 [#]	72.8	70.5	-2.4
Qwen 3.6 [#]	69.5	66.6	-2.8
Gemma 4	70.1	67.2	-2.9
Qwen 3.6	57.4	57.2	-0.3
xCOMET XL	52.7	58.4	+5.7
xCOMET XXL	56.3	61.1	+4.8

Table 62: Scores behavior of strictly matched reference-free and reference-based evaluators on human-evaluated translations. Values are averages over the eight priority directions; Δ score is reference-based minus reference-free. [#] marks reasoning auto-evaluators. Light-grey rows are organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4 - thinking	Gemma 4 31B IT	36.80	29.36	32.66
2	1	PhraseT1-01*	Gemini 3.5 Flash	37.21	28.54	32.30
3	1	Gemma 4	Gemma 4 31B IT	31.50	31.84	31.67
4	1	Gemma 4*	Gemma 4 31B IT	35.21	28.71	31.63
5	2	Gemma 4 - thinking*	Gemma 4 31B IT	38.62	26.49	31.43
6	2	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	44.87	23.97	31.24
7	2	fluencyscore-multispan-v2*	Gemini 3.5 Flash	44.94	23.93	31.23
8	2	Cohere CAT+*	Command A+	31.58	30.78	31.17
9	2	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	30.03	32.27	31.11
10	2	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	30.90	31.16	31.03
11	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	44.09	23.63	30.77
12	2	PhraseT1-02*	Ensemble ^{E1}	26.17	36.76	30.58
13	2	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	35.83	26.37	30.38
14	2	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	26.91	34.45	30.21
15	3	GPT-5.6 Terra*	GPT-5.6 Terra	32.54	27.40	29.75
16	3	cuni-cat-v4*	Gemma 4 31B	40.07	23.42	29.56
17	3	cuni-v14*	Gemma 4 31B	38.73	23.19	29.01
18	4	Qwen 3.6*	Qwen3.6-35B-A3B	18.71	37.23	24.90
19	4	Qwen 3.6	Qwen3.6-35B-A3B	17.70	40.48	24.63
20	4	doc-FACET*	GPT-4.1 Mini (2025-04-14)	25.17	23.73	24.43
21	4	FACET*	GPT-4.1 Mini (2025-04-14)	25.16	23.72	24.42
22	5	xCOMET XXL [†]	xCOMET-XXL	16.13	24.09	19.32
23	6	xCOMET XL [†]	xCOMET-XL	12.35	31.54	17.76
24	7	Qwen-based Omission ^{*†}	Qwen3-32B	15.33	11.55	13.17
25	7	Rule-based Omission ^{*†}	Qwen3-32B	15.27	11.48	13.11
26	8	xCOMET XL ^{*†}	xCOMET-XL	7.43	25.41	11.49
27	9	XCOMOffset [*]	Ensemble ^{E2}	6.78	17.72	9.81
28	9	xCOMET XXL ^{*†}	xCOMET-XXL	6.79	13.22	8.97

Table 18: Task 1 ranking for the Czech→Vietnamese translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [xcomet].

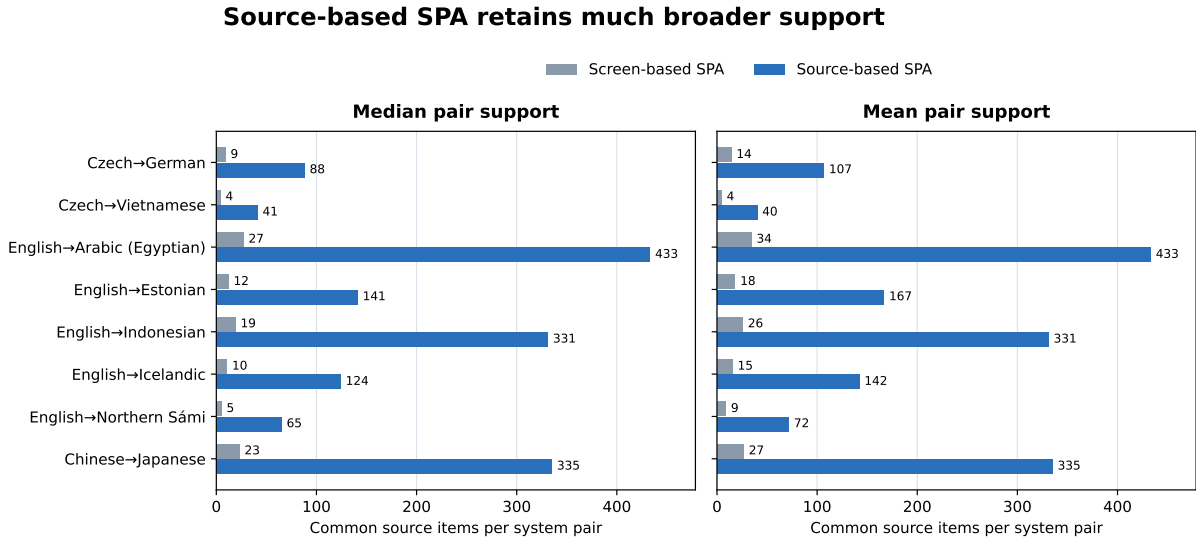


Figure 6: Human evidence support for the two system-level SPA variants across the eight primary translation directions. Each bar summarizes the number of source items with human scores for an MT system pair; the left and right panels report the median and mean, respectively, across system pairs within a direction. Screen-based SPA retains only source items where the two systems were assessed together in a cESA screen, whereas source-based SPA retains every source item with a human score for both systems.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	fluencyscore-multispan-v2*	Gemini 3.5 Flash	20.10	44.65	27.72
2	1	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	20.10	44.65	27.72
3	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	19.29	45.78	27.14
4	2	cuni-cat-v4*	Gemma 4 31B	16.68	52.35	25.30
5	2	cuni-v14*	Gemma 4 31B	16.55	52.39	25.15
6	3	GPT-5.6 Terra*	GPT-5.6 Terra	15.74	46.88	23.56
7	3	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	14.80	54.25	23.26
8	4	Gemma 4 - thinking*	Gemma 4 31B IT	14.69	51.25	22.83
9	5	Cohere CAT+	Command A+	14.77	42.66	21.94
10	5	Gemma 4 - thinking	Gemma 4 31B IT	13.38	55.58	21.56
11	6	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	13.27	51.65	21.11
12	6	Gemma 4*	Gemma 4 31B IT	12.73	55.88	20.74
13	7	FACET*	GPT-4.1 Mini (2025-04-14)	13.89	39.33	20.53
14	8	doc-FACET*	GPT-4.1 Mini (2025-04-14)	13.83	39.40	20.48
15	9	PhraseT1-01*	Gemini 3.5 Flash	12.35	48.08	19.66
16	9	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E1}	11.59	53.88	19.07
17	10	Gemma 4	Gemma 4 31B IT	10.90	61.58	18.52
18	11	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	11.06	44.34	17.71
19	12	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	9.60	50.19	16.12
20	13	PhraseT1-02*	Ensemble ^{E2}	8.84	58.22	15.35
21	14	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	8.95	37.73	14.47
22	15	Qwen 3.6*	Qwen3.6-35B-A3B	6.62	54.43	11.80
23	16	Qwen-based Omission* [†]	Qwen3-32B	7.10	18.59	10.27
24	17	Rule-based Omission* [†]	Qwen3-32B	7.07	18.42	10.22
25	17	Qwen 3.6	Qwen3.6-35B-A3B	5.57	60.67	10.20
26	18	xCOMET XXL [†]	xCOMET-XXL	4.74	44.19	8.56
27	19	xCOMET XXL* [†]	xCOMET-XXL	4.25	38.87	7.66
28	20	XCOMOffset*	Ensemble ^{E4}	3.60	19.93	6.10
29	21	xCOMET XL [†]	xCOMET-XL	2.93	44.73	5.50
30	22	xCOMET XL* [†]	xCOMET-XL	2.48	44.61	4.70

Table 19: Task 1 ranking for the English→Egyptian Arabic translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E2 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E4 [xcomet].

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4	Gemma 4 31B IT	39.83	37.03	38.38
2	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	56.81	27.85	37.38
3	2	Gemma 4 - thinking	Gemma 4 31B IT	47.63	30.65	37.30
4	3	Gemma 4*	Gemma 4 31B IT	42.46	30.94	35.80
5	3	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	56.76	25.74	35.42
6	3	fluencyscore-multispan-v2*	Gemini 3.5 Flash	56.90	25.71	35.42
7	4	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	34.32	35.39	34.85
8	4	Gemma 4 - thinking*	Gemma 4 31B IT	49.41	26.67	34.65
9	5	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	35.72	31.60	33.54
10	6	Qwen 3.6	Qwen3.6-35B-A3B	22.94	48.92	31.23
11	7	FACET*	GPT-4.1 Mini (2025-04-14)	34.78	26.42	30.03
12	7	doc-FACET*	GPT-4.1 Mini (2025-04-14)	34.69	26.45	30.01
13	7	Qwen 3.6*	Qwen3.6-35B-A3B	23.81	39.30	29.66
14	8	Cohere CAT+	Command A+	40.87	20.92	27.67
15	9	xCOMET XXL [†]	xCOMET-XXL	19.09	41.72	26.19
16	10	xCOMET XXL* [†]	xCOMET-XXL	16.84	39.92	23.69
17	11	xCOMET XL [†]	xCOMET-XL	16.09	41.32	23.16
18	12	xCOMET XL* [†]	xCOMET-XL	13.74	41.59	20.65

Table 20: Task 1 ranking for the English→Belarusian translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4 - thinking	Gemma 4 31B IT	34.94	31.51	33.14
2	1	Gemma 4	Gemma 4 31B IT	28.48	39.24	33.01
3	2	Gemma 4*	Gemma 4 31B IT	35.01	28.44	31.38
4	2	fluencyscore-multispan-v2*	Gemini 3.5 Flash	47.02	23.30	31.16
5	2	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	46.75	23.36	31.15
6	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	44.70	23.39	30.71
7	3	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	25.52	36.82	30.15
8	3	Gemma 4 - thinking*	Gemma 4 31B IT	39.23	24.44	30.11
9	3	Cohere CAT+*	Command A+	36.23	25.23	29.74
10	3	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	28.96	30.55	29.73
11	4	cuni-v14*	Gemma 4 31B	37.76	23.55	29.01
12	4	cuni-cat-v4*	Gemma 4 31B	37.81	23.23	28.78
13	5	Qwen 3.6*	Qwen3.6-35B-A3B	19.09	38.64	25.56
14	6	FACET*	GPT-4.1 Mini (2025-04-14)	24.57	22.16	23.30
15	7	doc-FACET*	GPT-4.1 Mini (2025-04-14)	24.56	22.16	23.29
16	7	Qwen 3.6	Qwen3.6-35B-A3B	14.97	49.65	23.00
17	8	xCOMET XXL [†]	xCOMET-XXL	14.98	37.32	21.38
18	9	xCOMET XXL ^{*†}	xCOMET-XXL	13.55	34.19	19.40
19	10	xCOMET XL [†]	xCOMET-XL	11.82	39.34	18.18
20	11	xCOMET XL ^{*†}	xCOMET-XL	9.66	39.58	15.53

Table 21: Task 1 ranking for the English→Czech translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4 - thinking	Gemma 4 31B IT	34.31	27.20	30.34
2	1	Gemma 4	Gemma 4 31B IT	27.47	33.76	30.29
3	2	Gemma 4*	Gemma 4 31B IT	34.07	25.84	29.39
4	3	Gemma 4 - thinking*	Gemma 4 31B IT	38.20	22.30	28.16
5	3	fluencyscore-multispan-v2*	Gemini 3.5 Flash	44.73	20.25	27.88
6	3	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	44.40	20.31	27.87
7	3	Cohere CAT+*	Command A+	31.22	25.01	27.77
8	4	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	23.60	31.99	27.16
9	4	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	27.89	25.86	26.83
10	4	Gemini-3.6-Flash*	Gemini 3.6 Flash	43.48	19.18	26.62
11	5	Qwen 3.6*	Qwen3.6-35B-A3B	19.68	33.82	24.89
12	6	Qwen 3.6	Qwen3.6-35B-A3B	15.50	43.99	22.92
13	6	FACET*	GPT-4.1 Mini (2025-04-14)	25.85	20.03	22.57
14	6	doc-FACET*	GPT-4.1 Mini (2025-04-14)	25.84	20.03	22.57
15	7	xCOMET XXL [†]	xCOMET-XXL	21.84	21.96	21.90
16	8	xCOMET XXL ^{*†}	xCOMET-XXL	19.54	20.09	19.81
17	9	QwenSpan ^{*†}	Qwen3-14B	30.93	12.97	18.28
18	10	xCOMET XL [†]	xCOMET-XL	13.01	21.22	16.13
19	11	xCOMET XL ^{*†}	xCOMET-XL	11.45	21.59	14.96

Table 22: Task 1 ranking for the English→German translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	PhraseT1-01*	Gemini 3.5 Flash	42.50	35.65	38.78
2	2	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	48.68	28.85	36.23
3	2	fluencyscore-multispan-v2*	Gemini 3.5 Flash	48.80	28.80	36.22
4	2	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	34.12	38.57	36.21
5	2	Gemma 4	Gemma 4 31B IT	32.88	39.46	35.87
6	3	Gemini-3.6-Flash*	Gemini 3.6 Flash	45.15	29.02	35.33
7	4	PhraseT1-02*	Ensemble ^{E1}	27.62	45.27	34.30
8	4	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	39.25	30.29	34.19
9	4	Gemma 4*	Gemma 4 31B IT	34.23	33.20	33.71
10	4	GPT-5.6 Terra*	GPT-5.6 Terra	37.18	30.70	33.63
11	5	Gemma 4 - thinking	Gemma 4 31B IT	35.68	31.40	33.40
12	6	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	27.22	39.54	32.25
13	7	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	28.77	34.32	31.30
14	7	cuni-v14*	Gemma 4 31B	35.48	27.59	31.04
15	7	cuni-cat-v4*	Gemma 4 31B	35.69	27.25	30.91
16	7	Gemma 4 - thinking*	Gemma 4 31B IT	36.35	26.54	30.68
17	8	Qwen 3.6*	Qwen3.6-35B-A3B	21.08	45.61	28.84
18	9	Qwen 3.6	Qwen3.6-35B-A3B	19.00	51.84	27.81
19	10	FACET*	GPT-4.1 Mini (2025-04-14)	28.92	25.35	27.02
20	11	doc-FACET*	GPT-4.1 Mini (2025-04-14)	28.90	25.36	27.02
21	12	xCOMET XXL [†]	xCOMET-XXL	19.48	38.38	25.84
22	12	Cohere CAT+*	Command A+	32.43	20.44	25.07
23	13	xCOMET XXL ^{*†}	xCOMET-XXL	17.76	36.30	23.85
24	14	xCOMET XL [†]	xCOMET-XL	15.67	42.28	22.87
25	15	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E2}	29.41	16.21	20.90
26	16	xCOMET XL ^{*†}	xCOMET-XL	13.15	41.15	19.93
27	17	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	27.24	14.29	18.74
28	18	XCOMOffset*	Ensemble ^{E4}	12.69	26.35	17.13
29	19	Qwen-based Omission ^{*†}	Qwen3-32B	21.69	10.70	14.33
30	20	Rule-based Omission ^{*†}	Qwen3-32B	21.67	10.64	14.27

Table 23: Task 1 ranking for the English→Estonian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E4 [xcomet].

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4	Gemma 4 31B IT	34.96	36.66	35.79
2	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	47.00	27.73	34.88
3	3	Gemma 4*	Gemma 4 31B IT	37.01	31.88	34.25
4	3	Gemma 4 - thinking	Gemma 4 31B IT	38.53	30.53	34.06
5	4	fluencyscore-multispan-v2*	Gemini 3.5 Flash	49.77	24.91	33.20
6	4	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	49.77	24.91	33.20
7	4	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	30.47	35.86	32.94
8	5	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	32.03	32.14	32.09
9	5	Gemma 4 - thinking*	Gemma 4 31B IT	39.65	26.38	31.68
10	6	Qwen 3.6*	Qwen3.6-35B-A3B	21.59	45.58	29.30
11	7	FACET*	GPT-4.1 Mini (2025-04-14)	31.04	24.73	27.52
12	7	doc-FACET*	GPT-4.1 Mini (2025-04-14)	31.04	24.73	27.52
13	7	Qwen 3.6	Qwen3.6-35B-A3B	22.33	35.00	27.26
14	8	xCOMET XXL [†]	xCOMET-XXL	20.23	37.38	26.25
15	9	Cohere CAT+*	Command A+	28.20	21.02	24.09
16	9	xCOMET XXL ^{*†}	xCOMET-XXL	17.54	37.01	23.80
17	9	xCOMET XL [†]	xCOMET-XL	17.73	35.47	23.64
18	10	xCOMET XL ^{*†}	xCOMET-XL	14.96	35.75	21.09

Table 24: Task 1 ranking for the English→Armenian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	fluencyscore-multispan-v2*	Gemini 3.5 Flash	33.81	29.44	31.48
2	1	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	33.72	29.49	31.46
3	1	PhraseT1-01*	Gemini 3.5 Flash	29.90	32.47	31.13
4	1	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	29.83	32.55	31.13
5	2	cuni-cat-v4*	Gemma 4 31B	30.19	31.38	30.77
6	2	Gemma 4 - thinking*	Gemma 4 31B IT	28.08	33.15	30.41
7	3	cuni-v14*	Gemma 4 31B	29.52	31.07	30.27
8	3	Gemini-3.6-Flash*	Gemini 3.6 Flash	33.17	27.74	30.21
9	3	Gemma 4 - thinking	Gemma 4 31B IT	25.96	35.47	29.98
10	3	Gemma 4*	Gemma 4 31B IT	25.04	36.73	29.78
11	4	Gemma 4	Gemma 4 31B IT	21.88	41.02	28.53
12	5	GPT-5.6 Terra*	GPT-5.6 Terra	24.92	30.75	27.53
13	5	PragmaSpan	GPT-5.5	20.21	40.91	27.05
14	6	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	21.43	36.34	26.96
15	6	PhraseT1-02*	Ensemble ^{E1}	20.10	40.84	26.94
16	6	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E2}	20.42	38.72	26.74
17	7	Cohere CAT+*	Command A+	21.88	32.49	26.15
18	8	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	19.15	37.00	25.24
19	9	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	17.48	40.09	24.35
20	10	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	17.60	36.29	23.70
21	11	FACET*	GPT-4.1 Mini (2025-04-14)	18.84	27.07	22.22
22	12	doc-FACET*	GPT-4.1 Mini (2025-04-14)	18.83	27.07	22.21
23	13	Qwen 3.6*	Qwen3.6-35B-A3B	12.76	43.86	19.76
24	14	Qwen 3.6	Qwen3.6-35B-A3B	11.48	47.65	18.51
25	15	xCOMET XXL [†]	xCOMET-XXL	11.61	35.49	17.50
26	15	Rule-based Omission ^{*†}	Qwen3-32B	16.64	17.11	16.87
27	15	Qwen-based Omission ^{*†}	Qwen3-32B	16.59	17.13	16.86
28	16	xCOMET XXL ^{*†}	xCOMET-XXL	10.58	35.39	16.28
29	17	xCOMET XL [†]	xCOMET-XL	7.84	42.25	13.23
30	18	xCOMET XL ^{*†}	xCOMET-XL	6.24	44.66	10.95

Table 25: Task 1 ranking for the English→Indonesian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus].

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	PhraseT1-01*	Gemini 3.5 Flash	42.93	42.77	42.85
2	2	Gemma 4 - thinking	Gemma 4 31B IT	41.20	40.61	40.91
3	2	Gemma 4	Gemma 4 31B IT	35.79	47.65	40.88
4	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	47.54	35.62	40.73
5	2	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	44.13	37.18	40.36
6	3	Gemma 4*	Gemma 4 31B IT	38.46	41.79	40.06
7	4	Gemma 4 - thinking*	Gemma 4 31B IT	42.39	35.89	38.87
8	4	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	34.29	44.11	38.58
9	4	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	49.03	31.74	38.53
10	4	fluencyscore-multispan-v2*	Gemini 3.5 Flash	49.13	31.68	38.52
11	5	cuni-cat-v4*	Gemma 4 31B	42.70	34.44	38.13
12	5	cuni-v14*	Gemma 4 31B	41.98	34.64	37.96
13	5	PhraseT1-02*	Ensemble ^{E1}	30.08	50.48	37.70
14	5	GPT-5.6 Terra*	GPT-5.6 Terra	37.85	37.21	37.53
15	6	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	30.10	43.66	35.63
16	6	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	31.88	39.42	35.25
17	7	FACET*	GPT-4.1 Mini (2025-04-14)	29.96	35.05	32.30
18	8	doc-FACET*	GPT-4.1 Mini (2025-04-14)	29.95	35.05	32.30
19	9	Qwen 3.6*	Qwen3.6-35B-A3B	22.72	49.46	31.14
20	9	Cohere CAT+*	Command A+	35.47	26.60	30.40
21	10	Qwen 3.6	Qwen3.6-35B-A3B	20.32	56.20	29.85
22	11	xCOMET XXL [†]	xCOMET-XXL	20.50	40.34	27.19
23	12	xCOMET XXL ^{*†}	xCOMET-XXL	18.69	38.47	25.16
24	13	xCOMET XL [†]	xCOMET-XL	16.08	44.64	23.64
25	14	xCOMET XL ^{*†}	xCOMET-XL	14.81	42.84	22.01
26	15	Qwen-based Omission ^{*†}	Qwen3-32B	21.60	9.93	13.61
27	15	Rule-based Omission ^{*†}	Qwen3-32B	21.60	9.92	13.59

Table 26: Task 1 ranking for the English→Icelandic translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B].

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	36.80	33.53	35.08
2	1	fluencyscore-multispan-v2*	Gemini 3.5 Flash	37.83	32.40	34.90
3	1	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	37.83	32.40	34.90
4	2	Gemma 4 - thinking	Gemma 4 31B IT	29.71	37.87	33.29
5	2	Gemma 4 - thinking*	Gemma 4 31B IT	32.12	33.68	32.88
6	3	Gemma 4	Gemma 4 31B IT	25.72	45.03	32.74
7	3	Gemma 4*	Gemma 4 31B IT	28.18	38.96	32.71
8	4	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	22.72	40.57	29.13
9	4	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	21.02	45.65	28.78
10	5	doc-FACET*	GPT-4.1 Mini (2025-04-14)	22.21	30.41	25.67
11	5	FACET*	GPT-4.1 Mini (2025-04-14)	22.21	30.41	25.67
12	6	Qwen 3.6*	Qwen3.6-35B-A3B	15.11	50.65	23.28
13	6	Cohere CAT+*	Command A+	24.28	22.34	23.27
14	7	Qwen 3.6	Qwen3.6-35B-A3B	14.00	56.72	22.45
15	8	xCOMET XXL [†]	xCOMET-XXL	12.19	37.64	18.42
16	8	xCOMET XL [†]	xCOMET-XL	11.78	38.51	18.05
17	9	xCOMET XXL ^{*†}	xCOMET-XXL	10.62	34.97	16.29
18	10	xCOMET XL ^{*†}	xCOMET-XL	9.58	38.84	15.36

Table 27: Task 1 ranking for the English→Kazakh translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Qwen 3.6	Qwen3.6-35B-A3B	15.23	36.29	21.46
2	2	Qwen 3.6*	Qwen3.6-35B-A3B	16.26	27.91	20.55
3	3	xCOMET XL [†]	xCOMET-XL	11.90	40.82	18.42
4	3	Gemma 4	Gemma 4 31B IT	24.17	14.82	18.37
5	3	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	21.09	16.08	18.24
6	4	xCOMET XXL [†]	xCOMET-XXL	12.04	31.89	17.48
7	5	xCOMET XL ^{*†}	xCOMET-XL	11.18	29.02	16.14
8	6	Gemma 4 - thinking	Gemma 4 31B IT	25.00	11.16	15.43
9	6	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	21.86	11.54	15.10
10	7	xCOMET XXL ^{*†}	xCOMET-XXL	11.96	19.34	14.78
11	8	Gemma 4*	Gemma 4 31B IT	23.58	9.90	13.95
12	9	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	30.92	8.52	13.36
13	10	fluencyscore-multispan-v2*	Gemini 3.5 Flash	30.91	8.51	13.35
14	10	FACET*	GPT-4.1 Mini (2025-04-14)	23.00	9.06	12.99
15	10	doc-FACET*	GPT-4.1 Mini (2025-04-14)	22.97	9.06	12.99
16	11	Gemma 4 - thinking*	Gemma 4 31B IT	25.27	7.62	11.71
17	11	Gemini-3.6-Flash*	Gemini 3.6 Flash	31.07	7.05	11.50
18	12	Cohere CAT+*	Command A+	24.56	6.98	10.87

Table 28: Task 1 ranking for the English→Ligurian translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4	Gemma 4 31B IT	27.21	24.43	25.75
2	2	Qwen 3.6*	Qwen3.6-35B-A3B	18.98	35.31	24.69
3	2	Qwen 3.6	Qwen3.6-35B-A3B	16.50	45.85	24.27
4	3	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	25.08	20.99	22.85
5	4	xCOMET XL [†]	xCOMET-XL	14.29	42.81	21.43
6	4	Gemma 4*	Gemma 4 31B IT	26.89	17.78	21.40
7	4	Gemma 4 - thinking	Gemma 4 31B IT	29.68	16.37	21.10
8	5	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	25.69	15.07	19.00
9	5	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	39.62	12.17	18.62
10	6	fluencyscore-multispan-v2*	Gemini 3.5 Flash	39.63	12.16	18.61
11	7	xCOMET XL ^{*†}	xCOMET-XL	13.15	26.09	17.48
12	7	FACET*	GPT-4.1 Mini (2025-04-14)	22.96	13.81	17.24
13	8	doc-FACET*	GPT-4.1 Mini (2025-04-14)	22.95	13.81	17.24
14	8	Gemma 4 - thinking*	Gemma 4 31B IT	29.63	11.52	16.59
15	9	Gemini-3.6-Flash*	Gemini 3.6 Flash	39.46	9.88	15.80
16	10	Cohere CAT+*	Command A+	25.61	9.64	14.01
17	10	xCOMET XXL ^{*†}	xCOMET-XXL	19.37	10.25	13.41
18	11	xCOMET XXL [†]	xCOMET-XXL	17.05	10.51	13.01

Table 29: Task 1 ranking for the English→Ladin translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4	Gemma 4 31B IT	27.82	36.38	31.52
2	2	Gemma 4 - thinking	Gemma 4 31B IT	32.64	29.24	30.84
3	2	fluencyscore-multispan-v2*	Gemini 3.5 Flash	39.35	25.15	30.69
4	2	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	39.12	25.24	30.69
5	2	Gemma 4*	Gemma 4 31B IT	30.03	31.09	30.55
6	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	38.73	25.10	30.46
7	3	Gemma 4 - thinking*	Gemma 4 31B IT	32.87	26.56	29.38
8	4	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	22.68	36.13	27.86
9	4	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	24.53	31.24	27.48
10	5	Cohere CAT+*	Command A+	25.53	27.95	26.68
11	6	Qwen 3.6*	Qwen3.6-35B-A3B	16.47	39.83	23.31
12	7	doc-FACET*	GPT-4.1 Mini (2025-04-14)	21.46	23.17	22.28
13	7	FACET*	GPT-4.1 Mini (2025-04-14)	21.46	23.16	22.28
14	7	Qwen 3.6	Qwen3.6-35B-A3B	14.56	46.46	22.17
15	8	xCOMET XXL [†]	xCOMET-XXL	12.90	29.97	18.04
16	9	xCOMET XXL ^{*†}	xCOMET-XXL	11.95	28.93	16.91
17	10	xCOMET XL [†]	xCOMET-XL	10.24	31.62	15.48
18	11	xCOMET XL ^{*†}	xCOMET-XL	8.49	33.52	13.55
19	12	XCOMOffset*	Ensemble ^{E1}	8.11	30.92	12.85

Table 30: Task 1 ranking for the English→Russian translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [xcomet].

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	55.26	41.48	47.39
2	2	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	39.27	49.73	43.89
3	2	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	44.68	42.81	43.73
4	2	PhraseT1-01*	Gemini 3.5 Flash	52.69	37.04	43.50
5	2	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E1}	40.45	46.38	43.21
6	3	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	40.33	45.93	42.95
7	3	Gemma 4	Gemma 4 31B IT	35.73	52.75	42.60
8	3	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	48.39	37.49	42.25
9	4	Gemma 4 - thinking	Gemma 4 31B IT	39.01	44.78	41.70
10	4	GPT-5.6 Terra*	GPT-5.6 Terra	45.62	37.25	41.01
11	4	fluencyscore-multispan-v2*	Gemini 3.5 Flash	56.61	31.27	40.29
12	4	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	56.61	31.27	40.29
13	5	Gemma 4*	Gemma 4 31B IT	33.19	50.73	40.13
14	5	Qwen 3.6*	Qwen3.6-35B-A3B	30.44	55.57	39.34
15	6	Gemma 4 - thinking*	Gemma 4 31B IT	36.56	41.16	38.72
16	7	Qwen 3.6	Qwen3.6-35B-A3B	26.06	62.92	36.86
17	7	cuni-cat-v4*	Gemma 4 31B	38.22	33.20	35.54
18	8	cuni-v14*	Gemma 4 31B	37.92	32.83	35.19
19	8	PhraseT1-02*	Ensemble ^{E2}	27.69	46.64	34.75
20	9	Cohere CAT+*	Command A+	34.13	24.93	28.82
21	10	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	31.81	20.90	25.23
22	11	xCOMET XL [†]	xCOMET-XL	14.89	36.30	21.12
23	11	doc-FACET*	GPT-4.1 Mini (2025-04-14)	38.76	14.02	20.59
24	11	FACET*	GPT-4.1 Mini (2025-04-14)	38.76	14.01	20.58
25	11	xCOMET XL ^{*†}	xCOMET-XL	17.76	23.30	20.16
26	12	xCOMET XXL [†]	xCOMET-XXL	14.19	8.71	10.79
27	12	Qwen-based Omission ^{*†}	Qwen3-32B	29.13	6.31	10.37
28	12	Rule-based Omission ^{*†}	Qwen3-32B	29.08	6.29	10.34
29	13	xCOMET XXL ^{*†}	xCOMET-XXL	24.22	5.35	8.76

Table 31: Task 1 ranking for the English→Northern Sami translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E2 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus].

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4*	Gemma 4 31B IT	40.69	38.41	39.52
2	2	Gemma 4	Gemma 4 31B IT	35.92	42.10	38.77
3	3	Gemma 4 - thinking	Gemma 4 31B IT	40.58	35.64	37.95
4	4	Gemma 4 - thinking*	Gemma 4 31B IT	44.15	32.62	37.52
5	5	Gemini-3.6-Flash*	Gemini 3.6 Flash	49.69	28.71	36.39
6	6	fluencyscore-multispan-v2*	Gemini 3.5 Flash	48.07	26.63	34.27
7	6	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	48.07	26.63	34.27
8	6	Cohere CAT+*	Command A+	34.56	32.61	33.56
9	7	FACET*	GPT-4.1 Mini (2025-04-14)	33.05	27.09	29.77
10	7	doc-FACET*	GPT-4.1 Mini (2025-04-14)	33.05	27.09	29.77
11	8	xCOMET XXL [†]	xCOMET-XXL	19.71	33.92	24.93
12	9	xCOMET XXL ^{*†}	xCOMET-XXL	17.56	31.95	22.66
13	10	xCOMET XL [†]	xCOMET-XL	14.59	37.79	21.05
14	11	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	14.37	26.11	18.54
15	12	xCOMET XL ^{*†}	xCOMET-XL	11.67	38.02	17.85
16	12	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	13.86	24.18	17.62
17	13	Qwen 3.6*	Qwen3.6-35B-A3B	6.73	43.11	11.65
18	14	Qwen 3.6	Qwen3.6-35B-A3B	6.37	49.58	11.29

Table 32: Task 1 ranking for the English→Thai translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	fluencyscore-multispan-v2*	Gemini 3.5 Flash	37.02	32.38	34.55
2	2	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	36.88	32.43	34.51
3	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	36.74	32.25	34.35
4	2	Gemma 4 - thinking	Gemma 4 31B IT	30.95	38.24	34.21
5	3	Gemma 4 - thinking*	Gemma 4 31B IT	32.03	34.87	33.39
6	4	Gemma 4	Gemma 4 31B IT	25.41	45.75	32.68
7	5	Gemma 4*	Gemma 4 31B IT	26.93	39.16	31.91
8	6	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	23.11	38.09	28.77
9	6	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	21.53	42.70	28.63
10	7	Cohere CAT+*	Command A+	26.35	29.82	27.98
11	8	FACET*	GPT-4.1 Mini (2025-04-14)	21.51	29.50	24.88
12	9	doc-FACET*	GPT-4.1 Mini (2025-04-14)	21.50	29.50	24.88
13	10	Qwen 3.6*	Qwen3.6-35B-A3B	15.50	45.66	23.14
14	11	Qwen 3.6	Qwen3.6-35B-A3B	13.17	52.68	21.07
15	12	xCOMET XXL [†]	xCOMET-XXL	11.84	41.67	18.44
16	13	xCOMET XXL ^{*†}	xCOMET-XXL	11.07	39.87	17.33
17	14	xCOMET XL [†]	xCOMET-XL	9.08	42.59	14.96
18	15	xCOMET XL ^{*†}	xCOMET-XL	7.64	42.77	12.96
19	16	XCOMOffset*	Ensemble ^{E1}	6.73	34.67	11.27

Table 33: Task 1 ranking for the English→Ukrainian translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [xcomet].

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4	Gemma 4 31B IT	30.14	32.33	31.20
2	2	Gemma 4*	Gemma 4 31B IT	30.31	28.37	29.31
3	2	Gemma 4 - thinking	Gemma 4 31B IT	32.65	26.42	29.20
4	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	40.28	21.80	28.29
5	3	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	22.44	36.77	27.87
6	3	cuni-v14*	Gemma 4 31B	36.24	22.64	27.87
7	3	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	24.71	31.61	27.73
8	3	cuni-cat-v4*	Gemma 4 31B	35.79	22.58	27.69
9	4	Gemma 4 - thinking*	Gemma 4 31B IT	30.95	23.41	26.66
10	5	fluencyscore-multispan-v2*	Gemini 3.5 Flash	26.10	22.67	24.26
11	5	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	26.10	22.67	24.26
12	6	Qwen 3.6	Qwen3.6-35B-A3B	13.67	48.39	21.32
13	6	doc-FACET*	GPT-4.1 Mini (2025-04-14)	25.53	17.98	21.10
14	6	FACET*	GPT-4.1 Mini (2025-04-14)	25.53	17.98	21.10
15	6	Qwen 3.6*	Qwen3.6-35B-A3B	14.87	35.06	20.89
16	7	Cohere CAT+*	Command A+	18.37	22.51	20.23
17	8	xCOMET XXL [†]	xCOMET-XXL	13.27	31.64	18.70
18	8	QwenSpan ^{*†}	Qwen3-14B	20.45	15.77	17.81
19	9	xCOMET XXL ^{*†}	xCOMET-XXL	10.77	31.10	16.00
20	10	xCOMET XL [†]	xCOMET-XL	7.26	45.68	12.53
21	11	xCOMET XL ^{*†}	xCOMET-XL	6.08	47.43	10.78

Table 34: Task 1 ranking for the English→Chinese (Simplified) translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	Gemma 4	Gemma 4 31B IT	31.92	28.12	29.90
2	2	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	25.37	30.93	27.88
3	2	Gemma 4 - thinking	Gemma 4 31B IT	35.03	22.30	27.25
4	3	Gemma 4*	Gemma 4 31B IT	33.89	21.51	26.32
5	3	Gemini-3.6-Flash*	Gemini 3.6 Flash	43.53	18.18	25.65
6	4	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	25.67	24.70	25.18
7	4	cuni-cat-v4*	Gemma 4 31B	36.61	18.45	24.54
8	4	Gemma 4 - thinking*	Gemma 4 31B IT	33.50	19.24	24.44
9	4	cuni-v14*	Gemma 4 31B	36.12	18.41	24.39
10	5	fluencyscore-multispan-v2*	Gemini 3.5 Flash	29.25	20.15	23.86
11	5	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	29.25	20.15	23.86
12	5	Qwen 3.6*	Qwen3.6-35B-A3B	17.28	34.66	23.06
13	6	Qwen 3.6	Qwen3.6-35B-A3B	14.95	42.96	22.18
14	7	Cohere CAT+*	Command A+	19.91	17.67	18.72
15	7	xCOMET XXL [†]	xCOMET-XXL	13.64	28.59	18.47
16	7	doc-FACET*	GPT-4.1 Mini (2025-04-14)	26.72	13.85	18.25
17	7	FACET*	GPT-4.1 Mini (2025-04-14)	26.72	13.85	18.25
18	8	xCOMET XXL ^{*†}	xCOMET-XXL	12.16	29.91	17.29
19	9	xCOMET XL [†]	xCOMET-XL	8.14	38.21	13.43
20	10	xCOMET XL ^{*†}	xCOMET-XL	6.65	40.82	11.43

Table 35: Task 1 ranking for the English→Chinese (Traditional, Taiwan) translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows indicate organizer baselines.

Rank	Cluster	Auto-evaluator	Underlying model	P	R	F1
1	1	cuni-v14*	Gemma 4 31B	31.10	39.77	34.90
2	1	cuni-cat-v4*	Gemma 4 31B	31.09	39.63	34.84
3	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	28.67	38.93	33.03
4	2	fluencyscore-multispan-v2*	Gemini 3.5 Flash	28.46	38.29	32.65
5	3	fluencyscore-multispan-v2.1*	Gemini 3.5 Flash	28.39	38.33	32.62
6	3	Gemma 4 - thinking*	Gemma 4 31B IT	26.73	41.52	32.53
7	4	gemma4-31b-v4ref-struct-best	Gemma 4 31B IT	25.05	43.65	31.83
8	4	Gemma 4*	Gemma 4 31B IT	25.22	42.15	31.56
9	5	GPT-5.6 Terra*	GPT-5.6 Terra	23.49	43.28	30.45
10	6	Gemma 4 - thinking	Gemma 4 31B IT	21.97	46.77	29.90
11	7	PhraseT1-01*	Gemini 3.5 Flash	22.93	40.80	29.36
12	8	GPT-5.6 Sol - medium thinking*	GPT-5.6 Sol	19.57	49.44	28.04
13	8	Gemma 4	Gemma 4 31B IT	19.43	50.02	27.99
14	9	Qwen 3.6 - thinking*	Qwen3.6-35B-A3B	19.59	46.04	27.48
15	9	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E1}	18.77	50.35	27.35
16	10	PhraseT1-02*	Ensemble ^{E2}	18.29	46.96	26.33
17	11	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	17.48	44.52	25.10
18	11	Qwen 3.6 - thinking	Qwen3.6-35B-A3B	16.15	52.21	24.67
19	12	FACET*	GPT-4.1 Mini (2025-04-14)	19.37	29.98	23.53
20	13	doc-FACET*	GPT-4.1 Mini (2025-04-14)	19.29	30.04	23.49
21	13	Cohere CAT+*	Command A+	16.35	40.97	23.38
22	14	Qwen 3.6*	Qwen3.6-35B-A3B	12.25	49.69	19.66
23	14	Qwen-based Omission ^{*†}	Qwen3-32B	17.62	20.85	19.10
24	14	Rule-based Omission ^{*†}	Qwen3-32B	17.63	20.79	19.08
25	15	Qwen 3.6	Qwen3.6-35B-A3B	10.10	56.07	17.12
26	16	xCOMET XXL [†]	xCOMET-XXL	9.30	33.54	14.56
27	17	xCOMET XL [†]	xCOMET-XL	7.64	37.02	12.67
28	18	xCOMET XXL ^{*†}	xCOMET-XXL	6.97	22.36	10.63
29	19	xCOMET XL ^{*†}	xCOMET-XL	4.97	38.14	8.80
30	20	XCOMOffset*	Ensemble ^{E4}	3.98	21.70	6.73

Table 36: Task 1 ranking for the Chinese (Simplified)→Japanese translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows indicate organizer baselines. The ensemble components listed in the model card are E1 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E2 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]; E4 [xcomet].

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.652	0.902	0.777
2	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.634	0.894	0.764
3	2	GPT-5.6 Terra*	GPT-5.6 Terra	0.627	0.895	0.761
4	2	bytepop_pro9	Ensemble ^{E2}	0.625	0.892	0.758
5	2	bytepop_pro10	Ensemble ^{E2}	0.625	0.892	0.758
6	2	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.632	0.880	0.756
7	2	Pterodactyl*	GPT-5.4	0.612	0.896	0.754
8	2	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.612	0.896	0.754
9	3	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.613	0.880	0.746
10	3	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.582	0.908	0.745
11	3	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.619	0.872	0.745
12	3	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.609	0.876	0.742
13	4	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E3}	0.567	0.896	0.732
14	4	cuni-da-6d*	Gemma 4 31B IT	0.574	0.871	0.722
15	4	Cohere CAT+ ensemble*	Command A+	0.562	0.883	0.722
16	4	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.555	0.883	0.719
17	5	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	0.529	0.895	0.712
18	5	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.498	0.908	0.703
19	5	cuni-v14-regression*	Gemma 4 31B IT	0.540	0.857	0.699
20	5	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.510	0.887	0.698
21	5	PhraseT2-02*	Ensemble ^{E1}	0.490	0.904	0.697
22	6	chrF	–	0.471	0.892	0.681
23	6	spBLEU	–	0.472	0.881	0.677
24	6	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.477	0.859	0.668
25	7	BLEU	–	0.464	0.871	0.667
26	7	xCOMET XXL [†]	xCOMET-XXL	0.483	0.838	0.660
27	7	COMET22	XLM-RoBERTa-large	0.499	0.818	0.658
28	7	Vertical_2 [†]	Gemma 3 1B IT	0.453	0.844	0.649
29	7	xCOMET XL [†]	xCOMET-XL	0.467	0.824	0.646
30	8	AEGIS	XLM-RoBERTa-large; mT5-large	0.475	0.814	0.645
31	8	BERTScore	XLM-RoBERTa-large	0.427	0.841	0.634
32	8	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.388	0.868	0.628
33	8	MQM-LLM ^{*†}	Not specified (27B)	0.412	0.833	0.623
34	8	gemba-poly*	Llama 3.3 70B	0.387	0.857	0.622
35	8	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.417	0.822	0.619
36	9	Vertical_1 [†]	Gemma 3 1B IT	0.415	0.805	0.610
37	9	YiSi-1	XLM-RoBERTa-large	0.328	0.871	0.600
38	9	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.340	0.821	0.580
39	9	FACET*	GPT-4.1 Mini (2025-04-14)	0.340	0.818	0.579
40	10	xCOMET XXL ^{*†}	xCOMET-XXL	0.340	0.801	0.571
41	10	xCOMET XL ^{*†}	xCOMET-XL	0.343	0.778	0.560
42	10	COMETKiwi22*	XLM-RoBERTa-large	0.345	0.745	0.545

Table 37: Task 2 ranking on the Czech→German translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.679	0.919	0.799
2	2	bytepop_pro9	Ensemble ^{E2}	0.654	0.922	0.788
3	2	bytepop_pro10	Ensemble ^{E2}	0.654	0.922	0.788
4	2	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.657	0.918	0.787
5	2	Pterodactyl*	GPT-5.4	0.658	0.910	0.784
6	2	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.658	0.910	0.784
7	2	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.646	0.916	0.781
8	3	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.643	0.914	0.778
9	3	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.616	0.917	0.767
10	4	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.611	0.913	0.762
11	5	cuni-v14-regression*	Gemma 4 31B IT	0.579	0.912	0.746
12	5	Cohere CAT+ ensemble*	Command A+	0.568	0.911	0.739
13	5	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.562	0.898	0.730
14	6	MQM-LLM ^{*†}	Not specified (27B)	0.560	0.888	0.724
15	6	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.533	0.904	0.719
16	6	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.508	0.911	0.710
17	7	gemba-poly*	Llama 3.3 70B	0.476	0.905	0.690
18	8	AEGIS	XLm-RoBERTa-large; mT5-large	0.493	0.822	0.657
19	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.417	0.874	0.645
20	8	COMETKiwi22*	XLm-RoBERTa-large	0.452	0.833	0.643
21	8	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.380	0.904	0.642
22	9	xCOMET XL ^{*†}	xCOMET-XL	0.427	0.831	0.629
23	9	Vertical_1 [†]	Gemma 3 1B IT	0.458	0.782	0.620
24	9	FACET*	GPT-4.1 Mini (2025-04-14)	0.338	0.886	0.612
25	9	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.337	0.885	0.611
26	10	Vertical_2 [†]	Gemma 3 1B IT	0.403	0.796	0.600

Table 38: Task 2 ranking on the Czech→Ukrainian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Because no reference is available in this direction, the reference-based Gemma 4 reasoning, xCOMET XL, and xCOMET XXL baselines are omitted; their reference-free counterparts are displayed. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	GPT-5.6 Terra*	GPT-5.6 Terra	0.668	0.902	0.785
2	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.674	0.894	0.784
3	1	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.677	0.890	0.783
4	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.672	0.892	0.782
5	1	bytopop_pro9	Ensemble ^{E2}	0.658	0.894	0.776
6	1	bytopop_pro10	Ensemble ^{E2}	0.658	0.894	0.776
7	1	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.665	0.883	0.774
8	1	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.657	0.888	0.772
9	1	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.648	0.896	0.772
10	2	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.650	0.881	0.765
11	2	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.642	0.888	0.765
12	2	Pterodactyl*	GPT-5.4	0.639	0.888	0.764
13	2	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.639	0.888	0.764
14	2	cuni-v14-regression*	Gemma 4 31B IT	0.620	0.896	0.758
15	2	Cohere CAT+ ensemble*	Command A+	0.611	0.887	0.749
16	3	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.591	0.897	0.744
17	3	Vertical_2 [†]	Gemma 3 1B IT	0.593	0.873	0.733
18	3	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.564	0.869	0.716
19	3	COMET22	XLM-RoBERTa-large	0.573	0.859	0.716
20	4	chrF	–	0.575	0.857	0.716
21	4	MQM-LLM ^{*†}	Not specified (27B)	0.551	0.870	0.710
22	4	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.539	0.871	0.705
23	4	Vertical_1 [†]	Gemma 3 1B IT	0.557	0.852	0.704
24	4	gemba-poly*	Llama 3.3 70B	0.522	0.877	0.699
25	4	cuni-da-6d*	Gemma 4 31B IT	0.546	0.847	0.697
26	5	spBLEU	–	0.532	0.835	0.683
27	5	BLEU	–	0.521	0.835	0.678
28	5	AEGIS	XLM-RoBERTa-large; mT5-large	0.546	0.796	0.671
29	5	PhraseT2-02*	Ensemble ^{E1}	0.483	0.856	0.669
30	5	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.478	0.860	0.669
31	6	xCOMET XL [†]	xCOMET-XL	0.504	0.777	0.640
32	6	BERTScore	XLM-RoBERTa-large	0.426	0.845	0.636
33	6	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.426	0.837	0.631
34	6	xCOMET XXL [†]	xCOMET-XXL	0.483	0.774	0.628
35	6	FACET*	GPT-4.1 Mini (2025-04-14)	0.408	0.829	0.619
36	6	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.409	0.829	0.619
37	7	YiSi-1	XLM-RoBERTa-large	0.325	0.843	0.584
38	7	COMETKiwi22*	XLM-RoBERTa-large	0.358	0.715	0.536
39	8	xCOMET XL ^{*†}	xCOMET-XL	0.270	0.696	0.483
40	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.196	0.708	0.452

Table 39: Task 2 ranking on the Czech→Vietnamese translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.727	0.915	0.821
2	1	bytepop_pro9	Ensemble ^{E2}	0.741	0.896	0.819
3	1	bytepop_pro10	Ensemble ^{E2}	0.741	0.896	0.819
4	1	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.706	0.924	0.815
5	2	GPT-5.6 Terra*	GPT-5.6 Terra	0.706	0.912	0.809
6	3	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.665	0.910	0.788
7	3	Pterodactyl*	GPT-5.4	0.670	0.903	0.786
8	3	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.670	0.903	0.786
9	4	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.653	0.908	0.780
10	4	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.661	0.899	0.780
11	4	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.651	0.907	0.779
12	4	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.654	0.904	0.779
13	4	COMET22	XLM-RoBERTa-large	0.692	0.853	0.773
14	5	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.647	0.879	0.763
15	5	AEGIS	XLM-RoBERTa-large; mT5-large	0.678	0.818	0.748
16	6	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E3}	0.591	0.896	0.744
17	6	cuni-da-6d*	Gemma 4 31B IT	0.592	0.882	0.737
18	6	chrF	–	0.597	0.860	0.728
19	7	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.588	0.866	0.727
20	7	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.561	0.886	0.724
21	7	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	0.545	0.896	0.721
22	7	xCOMET XXL [†]	xCOMET-XXL	0.620	0.811	0.716
23	7	BERTScore	XLM-RoBERTa-large	0.583	0.848	0.715
24	7	Cohere CAT+ ensemble*	Command A+	0.544	0.863	0.704
25	8	gemba-poly*	Llama 3.3 70B	0.538	0.842	0.690
26	8	xCOMET XL [†]	xCOMET-XL	0.585	0.788	0.686
27	8	spBLEU	–	0.508	0.855	0.682
28	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.583	0.769	0.676
29	8	YiSi-1	XLM-RoBERTa-large	0.488	0.854	0.671
30	9	cuni-v14-regression*	Gemma 4 31B IT	0.511	0.810	0.660
31	9	BLEU	–	0.470	0.848	0.659
32	9	COMETKiwi22*	XLM-RoBERTa-large	0.602	0.683	0.642
33	10	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.351	0.869	0.610
34	10	FACET*	GPT-4.1 Mini (2025-04-14)	0.351	0.867	0.609
35	10	xCOMET XL ^{*†}	xCOMET-XL	0.496	0.689	0.592
36	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.294	0.883	0.589
37	11	Vertical_2 [†]	Gemma 3 1B IT	0.348	0.794	0.571
38	12	Vertical_1 [†]	Gemma 3 1B IT	0.324	0.779	0.552
39	12	PhraseT2-02*	Ensemble ^{E1}	0.283	0.792	0.537
40	12	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.325	0.749	0.537
41	13	MQM-LLM ^{*†}	Not specified (27B)	0.217	0.771	0.494
42	14	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.216	0.764	0.490

Table 40: Task 2 ranking on the English→Arabic (Egyptian) translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.757	0.927	0.842
2	1	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.751	0.926	0.839
3	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.728	0.937	0.833
4	2	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.730	0.921	0.825
5	2	bytepop_pro10	Ensemble ^{E2}	0.730	0.919	0.825
6	2	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.732	0.917	0.824
7	3	Pterodactyl†	GPT-5.4	0.729	0.908	0.818
8	3	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.729	0.908	0.818
9	3	bytepop_pro9	Ensemble ^{E2}	0.699	0.915	0.807
10	4	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.674	0.920	0.797
11	4	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.662	0.915	0.789
12	5	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.645	0.918	0.781
13	5	MQM-LLM ^{*†}	Not specified (27B)	0.629	0.908	0.768
14	5	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.624	0.910	0.767
15	6	xCOMET XXL [†]	xCOMET-XXL	0.633	0.886	0.759
16	6	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.598	0.908	0.753
17	6	Vertical_2 [†]	Gemma 3 1B IT	0.602	0.902	0.752
18	7	xCOMET XL [†]	xCOMET-XL	0.621	0.875	0.748
19	7	chrF	–	0.586	0.893	0.740
20	7	spBLEU	–	0.570	0.898	0.734
21	8	Vertical_1 [†]	Gemma 3 1B IT	0.544	0.900	0.722
22	8	COMET22	XLM-RoBERTa-large	0.576	0.864	0.720
23	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.562	0.862	0.712
24	8	BLEU	–	0.533	0.891	0.712
25	8	AEGIS	XLM-RoBERTa-large; mT5-large	0.561	0.845	0.703
26	9	gemba-poly*	Llama 3.3 70B	0.506	0.868	0.687
27	9	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.464	0.908	0.686
28	9	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.462	0.901	0.681
29	9	FACET*	GPT-4.1 Mini (2025-04-14)	0.462	0.899	0.681
30	9	Cohere CAT+ ensemble*	Command A+	0.480	0.875	0.677
31	9	xCOMET XL ^{*†}	xCOMET-XL	0.528	0.817	0.673
32	10	BERTScore	XLM-RoBERTa-large	0.426	0.872	0.649
33	10	COMETKiwi22*	XLM-RoBERTa-large	0.502	0.715	0.608
34	11	YiSi-1	XLM-RoBERTa-large	0.300	0.865	0.583

Table 41: Task 2 ranking on the English→Belarusian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Pterodactyl [*]	GPT-5.4	0.676	0.915	0.795
2	1	Austriadactylus [*]	GPT-5.4; GPT-5.6 Terra	0.676	0.915	0.795
3	1	bytepop_pro9	Ensemble ^{E2}	0.677	0.904	0.791
4	1	bytepop_pro10	Ensemble ^{E2}	0.677	0.904	0.791
5	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.674	0.906	0.790
6	1	Gemma 4 - Reasoning [*]	Gemma 4 31B IT	0.665	0.908	0.787
7	2	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.663	0.901	0.782
8	2	Gemma 4 - No Reasoning [*]	Gemma 4 31B IT	0.659	0.904	0.781
9	2	Cohere CAT+ ensemble [*]	Command A+	0.626	0.933	0.780
10	3	Qwen 3.6 - Reasoning [*]	Qwen3.6-35B-A3B	0.615	0.914	0.765
11	3	Gemini-3.6-Flash [*]	Gemini 3.6 Flash	0.634	0.892	0.763
12	3	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.600	0.908	0.754
13	3	spBLEU	–	0.590	0.891	0.740
14	4	xCOMET XXL [†]	xCOMET-XXL	0.610	0.865	0.737
15	4	cuni-v14-regression [*]	Gemma 4 31B IT	0.571	0.902	0.736
16	4	chrF	–	0.583	0.888	0.736
17	4	Qwen 3.6 - No Reasoning [*]	Qwen3.6-35B-A3B	0.559	0.907	0.733
18	4	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.562	0.902	0.732
19	5	BLEU	–	0.553	0.882	0.718
20	5	Vertical_2 [†]	Gemma 3 1B IT	0.549	0.879	0.714
21	5	xCOMET XL [†]	xCOMET-XL	0.558	0.860	0.709
22	5	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.506	0.903	0.705
23	5	MQM-LLM ^{*†}	Not specified (27B)	0.503	0.895	0.699
24	6	COMET22	XLM-RoBERTa-large	0.563	0.821	0.692
25	6	xCOMET XXL ^{*†}	xCOMET-XXL	0.529	0.848	0.689
26	6	AEGIS	XLM-RoBERTa-large; mT5-large	0.539	0.833	0.686
27	6	Vertical_1 [†]	Gemma 3 1B IT	0.494	0.850	0.672
28	6	fluency2-gemini35-esa [*]	Gemini 3.5 Flash	0.455	0.885	0.670
29	7	xCOMET XL ^{*†}	xCOMET-XL	0.484	0.819	0.652
30	7	BERTScore	XLM-RoBERTa-large	0.424	0.872	0.648
31	7	gemba-poly [*]	Llama 3.3 70B	0.407	0.871	0.639
32	7	COMETKiwi22 [*]	XLM-RoBERTa-large	0.498	0.779	0.638
33	8	FACET [*]	GPT-4.1 Mini (2025-04-14)	0.364	0.859	0.611
34	8	doc-FACET [*]	GPT-4.1 Mini (2025-04-14)	0.362	0.859	0.611
35	8	YiSi-1	XLM-RoBERTa-large	0.309	0.872	0.590

Table 42: Task 2 ranking on the English→Czech translation direction. ^{*} marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Pterodactyl [*]	GPT-5.4	0.666	0.917	0.792
2	1	Austriadactylus [*]	GPT-5.4; GPT-5.6 Terra	0.666	0.917	0.792
3	1	bytepop_pro9	Ensemble ^{E2}	0.670	0.913	0.791
4	1	bytepop_pro10	Ensemble ^{E2}	0.670	0.913	0.791
5	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.667	0.911	0.789
6	1	Gemma 4 - Reasoning [*]	Gemma 4 31B IT	0.656	0.919	0.788
7	1	Gemini-3.6-Flash [*]	Gemini 3.6 Flash	0.653	0.914	0.784
8	2	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.650	0.903	0.777
9	2	Gemma 4 - No Reasoning [*]	Gemma 4 31B IT	0.641	0.910	0.775
10	3	Cohere CAT+ ensemble [*]	Command A+	0.582	0.919	0.750
11	3	Qwen 3.6 - Reasoning [*]	Qwen3.6-35B-A3B	0.600	0.899	0.749
12	3	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.594	0.899	0.747
13	4	chrF	–	0.526	0.897	0.711
14	4	Qwen 3.6 - No Reasoning [*]	Qwen3.6-35B-A3B	0.532	0.886	0.709
15	4	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.513	0.892	0.703
16	4	MQM-LLM ^{*†}	Not specified (27B)	0.525	0.880	0.703
17	4	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.509	0.893	0.701
18	5	spBLEU	–	0.495	0.891	0.693
19	5	Vertical_2 [†]	Gemma 3 1B IT	0.501	0.851	0.676
20	6	COMET22	XLM-RoBERTa-large	0.528	0.821	0.674
21	6	BLEU	–	0.468	0.879	0.674
22	6	gemba-poly [*]	Llama 3.3 70B	0.462	0.863	0.662
23	6	BERTScore	XLM-RoBERTa-large	0.452	0.869	0.661
24	7	AEGIS	XLM-RoBERTa-large; mT5-large	0.508	0.782	0.645
25	7	xCOMET XXL [†]	xCOMET-XXL	0.509	0.779	0.644
26	7	fluency2-gemini35-esa [*]	Gemini 3.5 Flash	0.410	0.875	0.642
27	7	xCOMET XL [†]	xCOMET-XL	0.497	0.785	0.641
28	7	Vertical_1 [†]	Gemma 3 1B IT	0.470	0.804	0.637
29	8	COMETKiwi22 [*]	XLM-RoBERTa-large	0.472	0.752	0.612
30	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.462	0.753	0.608
31	8	YiSi-1	XLM-RoBERTa-large	0.340	0.869	0.605
32	8	FACET [*]	GPT-4.1 Mini (2025-04-14)	0.359	0.837	0.598
33	8	doc-FACET [*]	GPT-4.1 Mini (2025-04-14)	0.359	0.837	0.598
34	9	xCOMET XL ^{*†}	xCOMET-XL	0.434	0.738	0.586

Table 43: Task 2 ranking on the English→German translation direction. ^{*} marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	bytepop_pro10	Ensemble ^{E2}	0.782	0.963	0.872
2	2	bytepop_pro9	Ensemble ^{E2}	0.768	0.958	0.863
3	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.757	0.965	0.861
4	2	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.763	0.952	0.858
5	2	GPT-5.6 Terra*	GPT-5.6 Terra	0.754	0.957	0.855
6	2	Pterodactyl [†]	GPT-5.4	0.753	0.954	0.853
7	2	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.753	0.954	0.853
8	3	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.730	0.944	0.837
9	3	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.725	0.945	0.835
10	4	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.712	0.945	0.828
11	4	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.707	0.941	0.824
12	4	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.693	0.948	0.820
13	4	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.690	0.946	0.818
14	5	xCOMET XXL [†]	xCOMET-XXL	0.686	0.925	0.806
15	5	xCOMET XL [†]	xCOMET-XL	0.675	0.931	0.803
16	5	cuni-da-6d*	Gemma 4 31B IT	0.652	0.928	0.790
17	6	cuni-v14-regression*	Gemma 4 31B IT	0.644	0.933	0.788
18	6	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.622	0.950	0.786
19	6	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.626	0.945	0.785
20	6	Vertical_2 [†]	Gemma 3 1B IT	0.616	0.942	0.779
21	6	COMET22	XLM-RoBERTa-large	0.651	0.906	0.779
22	6	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.616	0.936	0.776
23	6	chrF	–	0.610	0.934	0.772
24	6	xCOMET XXL ^{*†}	xCOMET-XXL	0.645	0.896	0.770
25	7	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E3}	0.583	0.956	0.770
26	7	PhraseT2-02*	Ensemble ^{E1}	0.597	0.937	0.767
27	7	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.582	0.951	0.767
28	7	spBLEU	–	0.602	0.923	0.763
29	7	AEGIS	XLM-RoBERTa-large; mT5-large	0.633	0.881	0.757
30	7	Cohere CAT+ ensemble*	Command A+	0.600	0.905	0.753
31	7	xCOMET XL ^{*†}	xCOMET-XL	0.628	0.876	0.752
32	8	MQM-LLM ^{*†}	Not specified (27B)	0.562	0.942	0.752
33	8	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	0.552	0.944	0.748
34	8	Vertical_1 [†]	Gemma 3 1B IT	0.561	0.933	0.747
35	8	gemba-poly*	Llama 3.3 70B	0.554	0.928	0.741
36	8	BLEU	–	0.573	0.909	0.741
37	9	COMETKiwi22*	XLM-RoBERTa-large	0.574	0.845	0.710
38	9	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.494	0.921	0.708
39	10	FACET*	GPT-4.1 Mini (2025-04-14)	0.494	0.916	0.705
40	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.475	0.924	0.699
41	11	BERTScore	XLM-RoBERTa-large	0.462	0.871	0.667
42	12	YiSi-1	XLM-RoBERTa-large	0.319	0.884	0.601

Table 44: Task 2 ranking on the English→Estonian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.816	0.961	0.888
2	1	bytepop_pro9	Ensemble ^{E2}	0.818	0.954	0.886
3	1	bytepop_pro10	Ensemble ^{E2}	0.818	0.954	0.886
4	1	Pterodactyl*	GPT-5.4	0.805	0.964	0.885
5	1	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.805	0.964	0.885
6	2	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.806	0.951	0.878
7	2	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.804	0.950	0.877
8	3	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.791	0.955	0.873
9	3	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.791	0.952	0.871
10	3	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.784	0.952	0.868
11	4	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.772	0.954	0.863
12	4	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.761	0.956	0.859
13	5	AEGIS	XLm-RoBERTa-large; mT5-large	0.755	0.939	0.847
14	5	xCOMET XL [†]	xCOMET-XL	0.754	0.937	0.846
15	5	Vertical_2 [†]	Gemma 3 1B IT	0.757	0.933	0.845
16	5	MQM-LLM ^{*†}	Not specified (27B)	0.734	0.955	0.845
17	5	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.731	0.953	0.842
18	5	xCOMET XXL [†]	xCOMET-XXL	0.751	0.933	0.842
19	6	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.729	0.948	0.838
20	6	COMET22	XLm-RoBERTa-large	0.739	0.928	0.834
21	6	Vertical_1 [†]	Gemma 3 1B IT	0.729	0.933	0.831
22	6	chrF	–	0.729	0.926	0.827
23	7	xCOMET XXL ^{*†}	xCOMET-XXL	0.717	0.928	0.823
24	7	spBLEU	–	0.694	0.922	0.808
25	8	COMETKiwi22*	XLm-RoBERTa-large	0.712	0.898	0.805
26	8	xCOMET XL ^{*†}	xCOMET-XL	0.689	0.918	0.804
27	9	gemba-poly*	Llama 3.3 70B	0.599	0.923	0.761
28	9	BLEU	–	0.602	0.901	0.752
29	9	BERTScore	XLm-RoBERTa-large	0.591	0.897	0.744
30	10	Cohere CAT+ ensemble*	Command A+	0.576	0.871	0.723
31	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.463	0.938	0.701
32	11	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.446	0.920	0.683
33	11	FACET*	GPT-4.1 Mini (2025-04-14)	0.446	0.920	0.683
34	11	YiSi-1	XLm-RoBERTa-large	0.470	0.875	0.672

Table 45: Task 2 ranking on the English→Armenian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.656	0.902	0.779
2	1	GPT-5.6 Terra*	GPT-5.6 Terra	0.647	0.906	0.776
3	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.652	0.898	0.775
4	2	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.656	0.888	0.772
5	2	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.653	0.890	0.771
6	3	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.653	0.882	0.768
7	3	bytepop_pro9	Ensemble ^{E2}	0.644	0.878	0.761
8	3	bytepop_pro10	Ensemble ^{E2}	0.644	0.878	0.761
9	4	cuni-da-6d*	Gemma 4 31B IT	0.628	0.892	0.760
10	4	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.641	0.872	0.757
11	4	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.612	0.894	0.753
12	4	Pterodactyl*	GPT-5.4	0.624	0.874	0.749
13	4	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.624	0.874	0.749
14	5	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.606	0.888	0.747
15	5	cuni-v14-regression*	Gemma 4 31B IT	0.578	0.895	0.736
16	5	Cohere CAT+ ensemble*	Command A+	0.585	0.887	0.736
17	5	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.568	0.903	0.736
18	6	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E3}	0.590	0.877	0.733
19	6	Vertical_2 [†]	Gemma 3 1B IT	0.552	0.887	0.719
20	7	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	0.560	0.876	0.718
21	7	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.543	0.877	0.710
22	7	Vertical_1 [†]	Gemma 3 1B IT	0.537	0.873	0.705
23	7	COMETKiwi22*	XLM-RoBERTa-large	0.581	0.827	0.704
24	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.592	0.801	0.697
25	8	xCOMET XXL [†]	xCOMET-XXL	0.593	0.800	0.697
26	8	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.499	0.885	0.692
27	8	COMET22	XLM-RoBERTa-large	0.549	0.824	0.686
28	9	xCOMET XL [†]	xCOMET-XL	0.576	0.786	0.681
29	9	PhraseT2-02*	Ensemble ^{E1}	0.463	0.881	0.672
30	9	gemba-poly*	Llama 3.3 70B	0.466	0.876	0.671
31	10	xCOMET XL ^{*†}	xCOMET-XL	0.557	0.779	0.668
32	10	AEGIS	XLM-RoBERTa-large; mT5-large	0.555	0.774	0.664
33	10	chrF	–	0.467	0.838	0.653
34	11	spBLEU	–	0.417	0.856	0.636
35	12	BLEU	–	0.404	0.849	0.627
36	12	BERTScore	XLM-RoBERTa-large	0.434	0.813	0.624
37	13	YiSi-1	XLM-RoBERTa-large	0.372	0.840	0.606
38	14	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.303	0.865	0.584
39	15	MQM-LLM ^{*†}	Not specified (27B)	0.366	0.667	0.517
40	16	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.347	0.656	0.501
41	16	FACET*	GPT-4.1 Mini (2025-04-14)	0.269	0.691	0.480
42	17	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.269	0.690	0.479

Table 46: Task 2 ranking on the English→Indonesian translation direction. * marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	bytepop_pro10	Ensemble ^{E2}	0.851	0.937	0.894
2	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.836	0.940	0.888
3	2	bytepop_pro9	Ensemble ^{E2}	0.837	0.935	0.886
4	3	Pterodactyl*	GPT-5.4	0.822	0.927	0.874
5	3	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.822	0.927	0.874
6	3	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.825	0.922	0.874
7	3	GPT-5.6 Terra*	GPT-5.6 Terra	0.818	0.925	0.871
8	3	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.815	0.924	0.870
9	3	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.815	0.924	0.869
10	4	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.799	0.920	0.859
11	4	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.792	0.921	0.857
12	5	xCOMET XXL [†]	xCOMET-XXL	0.768	0.921	0.845
13	5	cuni-da-6d*	Gemma 4 31B IT	0.769	0.913	0.841
14	5	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.755	0.918	0.836
15	6	xCOMET XL [†]	xCOMET-XL	0.752	0.920	0.836
16	6	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.733	0.922	0.827
17	6	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.744	0.910	0.827
18	7	Vertical_2 [†]	Gemma 3 1B IT	0.727	0.913	0.820
19	7	cuni-v14-regression*	Gemma 4 31B IT	0.713	0.915	0.814
20	7	chrF	–	0.715	0.910	0.812
21	7	xCOMET XXL ^{*†}	xCOMET-XXL	0.721	0.900	0.810
22	7	spBLEU	–	0.705	0.913	0.809
23	7	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.680	0.938	0.809
24	7	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.700	0.914	0.807
25	7	COMET22	XLM-RoBERTa-large	0.718	0.893	0.806
26	8	AEGIS	XLM-RoBERTa-large; mT5-large	0.720	0.884	0.802
27	8	MQM-LLM ^{*†}	Not specified (27B)	0.672	0.915	0.794
28	8	PhraseT2-02*	Ensemble ^{E1}	0.658	0.926	0.792
29	8	xCOMET XL ^{*†}	xCOMET-XL	0.697	0.881	0.789
30	9	Vertical_1 [†]	Gemma 3 1B IT	0.675	0.897	0.786
31	9	BLEU	–	0.667	0.904	0.785
32	9	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.658	0.913	0.785
33	9	Cohere CAT+ ensemble*	Command A+	0.669	0.870	0.769
34	10	COMETKiwi22*	XLM-RoBERTa-large	0.654	0.824	0.739
35	10	gemba-poly*	Llama 3.3 70B	0.568	0.887	0.728
36	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.528	0.910	0.719
37	11	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.513	0.913	0.713
38	11	FACET*	GPT-4.1 Mini (2025-04-14)	0.513	0.912	0.712
39	11	BERTScore	XLM-RoBERTa-large	0.520	0.861	0.690
40	12	YiSi-1	XLM-RoBERTa-large	0.400	0.876	0.638

Table 47: Task 2 ranking on the English→Icelandic translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.749	0.914	0.831
2	2	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.743	0.912	0.828
3	2	Pterodactyl*	GPT-5.4	0.732	0.920	0.826
4	2	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.732	0.920	0.826
5	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.741	0.906	0.824
6	3	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.731	0.910	0.820
7	4	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.713	0.910	0.811
8	4	bytepop_pro9	Ensemble ^{E2}	0.709	0.906	0.807
9	4	bytepop_pro10	Ensemble ^{E2}	0.709	0.906	0.807
10	5	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.695	0.899	0.797
11	6	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.665	0.897	0.781
12	6	MQM-LLM ^{*†}	Not specified (27B)	0.645	0.913	0.779
13	6	xCOMET XL [†]	xCOMET-XL	0.652	0.905	0.779
14	6	Vertical_2 [†]	Gemma 3 1B IT	0.649	0.906	0.777
15	6	AEGIS	XLM-RoBERTa-large; mT5-large	0.634	0.918	0.776
16	6	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.628	0.918	0.773
17	6	xCOMET XXL [†]	xCOMET-XXL	0.623	0.923	0.773
18	6	COMET22	XLM-RoBERTa-large	0.624	0.917	0.771
19	6	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.636	0.903	0.769
20	6	chrF	–	0.630	0.905	0.768
21	7	Vertical_1 [†]	Gemma 3 1B IT	0.585	0.905	0.745
22	7	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.575	0.912	0.743
23	7	gemba-poly*	Llama 3.3 70B	0.583	0.887	0.735
24	8	spBLEU	–	0.555	0.889	0.722
25	8	BERTScore	XLM-RoBERTa-large	0.529	0.884	0.706
26	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.509	0.900	0.704
27	8	xCOMET XL ^{*†}	xCOMET-XL	0.541	0.850	0.695
28	9	COMETKiwi22*	XLM-RoBERTa-large	0.575	0.797	0.686
29	9	BLEU	–	0.479	0.872	0.676
30	9	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.444	0.891	0.668
31	9	FACET*	GPT-4.1 Mini (2025-04-14)	0.444	0.891	0.668
32	9	Cohere CAT+ ensemble*	Command A+	0.501	0.832	0.667
33	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.403	0.906	0.654
34	11	YiSi-1	XLM-RoBERTa-large	0.368	0.880	0.624

Table 48: Task 2 ranking on the English→Kazakh translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	spBLEU	–	0.466	0.858	0.662
2	2	chrF	–	0.422	0.869	0.645
3	3	BLEU	–	0.415	0.848	0.631
4	4	Vertical_2 [†]	Gemma 3 1B IT	0.345	0.863	0.604
5	5	bytetop_pro9	Ensemble ^{E2}	0.384	0.773	0.579
6	5	bytetop_pro10	Ensemble ^{E2}	0.384	0.773	0.579
7	5	BERTScore	XLm-RoBERTa-large	0.293	0.864	0.578
8	5	COMET22	XLm-RoBERTa-large	0.310	0.846	0.578
9	6	YiSi-1	XLm-RoBERTa-large	0.195	0.871	0.533
10	6	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.273	0.765	0.519
11	6	MQM-LLM ^{*†}	Not specified (27B)	0.273	0.762	0.517
12	6	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.280	0.751	0.515
13	7	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.279	0.733	0.506
14	7	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.265	0.736	0.501
15	7	Gemini-3.6-Flash [*]	Gemini 3.6 Flash	0.252	0.741	0.496
16	7	Vertical_1 [†]	Gemma 3 1B IT	0.203	0.783	0.493
17	8	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.211	0.757	0.484
18	8	Austriadactylus [*]	GPT-5.4; GPT-5.6 Terra	0.229	0.731	0.480
19	8	Gemma 4 - Reasoning [*]	Gemma 4 31B IT	0.214	0.735	0.474
20	9	Pterodactyl [*]	GPT-5.4	0.190	0.742	0.466
21	10	Gemma 4 - No Reasoning [*]	Gemma 4 31B IT	0.173	0.724	0.448
22	10	Qwen 3.6 - Reasoning [*]	Qwen3.6-35B-A3B	0.162	0.733	0.448
23	10	fluency2-gemini35-esa [*]	Gemini 3.5 Flash	0.135	0.746	0.441
24	11	gemba-poly [*]	Llama 3.3 70B	0.105	0.710	0.407
25	11	Cohere CAT+ ensemble [*]	Command A+	0.095	0.676	0.386
26	12	AEGIS	XLm-RoBERTa-large; mT5-large	0.113	0.629	0.371
27	13	FACET [*]	GPT-4.1 Mini (2025-04-14)	0.011	0.638	0.325
28	14	doc-FACET [*]	GPT-4.1 Mini (2025-04-14)	0.003	0.633	0.318
29	14	Qwen 3.6 - No Reasoning [*]	Qwen3.6-35B-A3B	0.046	0.575	0.310
30	15	xCOMET XL [†]	xCOMET-XL	−0.044	0.413	0.184
31	16	xCOMET XXL [†]	xCOMET-XXL	−0.111	0.324	0.106
32	17	xCOMET XL ^{*†}	xCOMET-XL	−0.154	0.303	0.075
33	18	xCOMET XXL ^{*†}	xCOMET-XXL	−0.207	0.265	0.029
34	19	COMETKiwi22 [*]	XLm-RoBERTa-large	−0.272	0.253	−0.010

Table 49: Task 2 ranking on the English→Ligurian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	spBLEU	–	0.597	0.863	0.730
2	2	chrF	–	0.565	0.854	0.709
3	2	BLEU	–	0.556	0.848	0.702
4	3	COMET22	XLM-RoBERTa-large	0.478	0.853	0.666
5	3	bytepop_pro9	Ensemble ^{E2}	0.495	0.799	0.647
6	3	bytepop_pro10	Ensemble ^{E2}	0.495	0.799	0.647
7	4	Gemini-3.6-Flash [*]	Gemini 3.6 Flash	0.454	0.806	0.630
8	4	Vertical_2 [†]	Gemma 3 1B IT	0.426	0.818	0.622
9	4	BERTScore	XLM-RoBERTa-large	0.401	0.841	0.621
10	5	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.397	0.791	0.594
11	5	Austriadactylus [*]	GPT-5.4; GPT-5.6 Terra	0.395	0.768	0.581
12	6	Pterodactyl [*]	GPT-5.4	0.377	0.772	0.574
13	6	YiSi-1	XLM-RoBERTa-large	0.280	0.838	0.559
14	6	fluency2-gemini35-esa [*]	Gemini 3.5 Flash	0.267	0.835	0.551
15	7	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.313	0.788	0.550
16	7	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.342	0.732	0.537
17	7	Vertical_1 [†]	Gemma 3 1B IT	0.283	0.787	0.535
18	7	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.327	0.725	0.526
19	8	Qwen 3.6 - Reasoning [*]	Qwen3.6-35B-A3B	0.259	0.752	0.505
20	8	AEGIS	XLM-RoBERTa-large; mT5-large	0.236	0.727	0.481
21	9	MQM-LLM ^{*†}	Not specified (27B)	0.147	0.772	0.460
22	9	Gemma 4 - Reasoning [*]	Gemma 4 31B IT	0.226	0.689	0.457
23	9	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.145	0.768	0.456
24	9	gemba-poly [*]	Llama 3.3 70B	0.180	0.716	0.448
25	10	Gemma 4 - No Reasoning [*]	Gemma 4 31B IT	0.174	0.682	0.428
26	11	FACET [*]	GPT-4.1 Mini (2025-04-14)	0.036	0.664	0.350
27	12	doc-FACET [*]	GPT-4.1 Mini (2025-04-14)	0.036	0.663	0.349
28	12	Qwen 3.6 - No Reasoning [*]	Qwen3.6-35B-A3B	0.061	0.627	0.344
29	12	xCOMET XXL [†]	xCOMET-XXL	−0.029	0.597	0.284
30	13	Cohere CAT+ ensemble [*]	Command A+	−0.029	0.485	0.228
31	14	xCOMET XL [†]	xCOMET-XL	−0.075	0.364	0.144
32	15	xCOMET XXL ^{*†}	xCOMET-XXL	−0.207	0.303	0.048
33	16	xCOMET XL ^{*†}	xCOMET-XL	−0.235	0.260	0.013
34	17	COMETKiwi22 [*]	XLM-RoBERTa-large	−0.304	0.214	−0.045

Table 50: Task 2 ranking on the English→Latin translation direction. ^{*} marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Pterodactyl [*]	GPT-5.4	0.697	0.945	0.821
2	1	Austriadactylus [*]	GPT-5.4; GPT-5.6 Terra	0.697	0.945	0.821
3	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.694	0.943	0.819
4	1	Gemini-3.6-Flash [*]	Gemini 3.6 Flash	0.690	0.946	0.818
5	1	bytepop_pro9	Ensemble ^{E2}	0.693	0.940	0.816
6	1	bytepop_pro10	Ensemble ^{E2}	0.693	0.940	0.816
7	2	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.692	0.937	0.814
8	2	Gemma 4 - Reasoning [*]	Gemma 4 31B IT	0.683	0.942	0.812
9	3	Gemma 4 - No Reasoning [*]	Gemma 4 31B IT	0.676	0.936	0.806
10	4	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.650	0.938	0.794
11	4	Qwen 3.6 - Reasoning [*]	Qwen3.6-35B-A3B	0.644	0.938	0.791
12	5	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.608	0.930	0.769
13	5	Cohere CAT+ ensemble [*]	Command A+	0.593	0.930	0.762
14	5	MQM-LLM ^{*†}	Not specified (27B)	0.598	0.917	0.757
15	6	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.590	0.917	0.753
16	6	Qwen 3.6 - No Reasoning [*]	Qwen3.6-35B-A3B	0.576	0.919	0.747
17	6	Vertical_2 [†]	Gemma 3 1B IT	0.579	0.894	0.736
18	7	Vertical_1 [†]	Gemma 3 1B IT	0.562	0.877	0.719
19	7	xCOMET XXL [†]	xCOMET-XXL	0.582	0.832	0.707
20	8	COMET22	XLM-RoBERTa-large	0.560	0.828	0.694
21	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.551	0.814	0.682
22	8	chrF	–	0.506	0.850	0.678
23	9	xCOMET XL [†]	xCOMET-XL	0.558	0.778	0.668
24	9	AEGIS	XLM-RoBERTa-large; mT5-large	0.565	0.768	0.667
25	9	BERTScore	XLM-RoBERTa-large	0.466	0.838	0.652
26	9	spBLEU	–	0.442	0.861	0.651
27	9	COMETKiwi22 [*]	XLM-RoBERTa-large	0.548	0.754	0.651
28	9	gemba-poly [*]	Llama 3.3 70B	0.437	0.863	0.650
29	10	xCOMET XL ^{*†}	xCOMET-XL	0.507	0.771	0.639
30	10	BLEU	–	0.405	0.848	0.627
31	11	YiSi-1	XLM-RoBERTa-large	0.391	0.839	0.615
32	11	fluency2-gemini35-esa [*]	Gemini 3.5 Flash	0.350	0.871	0.611
33	11	FACET [*]	GPT-4.1 Mini (2025-04-14)	0.335	0.856	0.595
34	11	doc-FACET [*]	GPT-4.1 Mini (2025-04-14)	0.334	0.856	0.595

Table 51: Task 2 ranking on the English→Russian translation direction. ^{*} marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	bytepop_pro9	Ensemble ^{E2}	0.864	0.956	0.910
2	1	bytepop_pro10	Ensemble ^{E2}	0.864	0.956	0.910
3	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.849	0.947	0.898
4	3	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.842	0.923	0.882
5	3	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.817	0.936	0.876
6	4	GPT-5.6 Terra*	GPT-5.6 Terra	0.815	0.921	0.868
7	4	Pterodactyl*	GPT-5.4	0.823	0.912	0.868
8	4	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.813	0.919	0.866
9	5	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.769	0.935	0.852
10	5	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.798	0.906	0.852
11	5	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.763	0.930	0.846
12	6	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.751	0.904	0.827
13	6	chrF	–	0.732	0.915	0.824
14	7	spBLEU	–	0.717	0.900	0.808
15	7	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.716	0.890	0.803
16	7	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.685	0.900	0.793
17	7	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.676	0.904	0.790
18	7	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E3}	0.711	0.861	0.786
19	7	MQM-LLM ^{*†}	Not specified (27B)	0.658	0.900	0.779
20	8	cuni-v14-regression*	Gemma 4 31B IT	0.651	0.879	0.765
21	8	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.608	0.919	0.764
22	8	Vertical_2 [†]	Gemma 3 1B IT	0.677	0.842	0.760
23	8	BLEU	–	0.616	0.860	0.738
24	9	BERTScore	XLM-RoBERTa-large	0.569	0.894	0.732
25	9	gemba-poly*	Llama 3.3 70B	0.573	0.890	0.731
26	9	cuni-da-6d*	Gemma 4 31B IT	0.603	0.836	0.720
27	9	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	0.621	0.816	0.719
28	9	COMET22	XLM-RoBERTa-large	0.561	0.858	0.710
29	10	Cohere CAT+ ensemble*	Command A+	0.602	0.796	0.699
30	10	Vertical_1 [†]	Gemma 3 1B IT	0.563	0.809	0.686
31	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.460	0.872	0.666
32	11	AEGIS	XLM-RoBERTa-large; mT5-large	0.504	0.787	0.646
33	11	YiSi-1	XLM-RoBERTa-large	0.380	0.888	0.634
34	12	FACET*	GPT-4.1 Mini (2025-04-14)	0.294	0.800	0.547
35	12	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.294	0.800	0.547
36	13	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.269	0.592	0.431
37	14	PhraseT2-02*	Ensemble ^{E1}	0.217	0.591	0.404
38	14	xCOMET XXL [†]	xCOMET-XXL	0.175	0.633	0.404
39	15	xCOMET XL [†]	xCOMET-XL	0.092	0.545	0.319
40	16	xCOMET XXL ^{*†}	xCOMET-XXL	−0.109	0.373	0.132
41	16	COMETKiwi22*	XLM-RoBERTa-large	−0.124	0.383	0.129
42	16	xCOMET XL ^{*†}	xCOMET-XL	−0.124	0.310	0.093

Table 52: Task 2 ranking on the English→Northern Sámi translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.820	0.968	0.894
2	1	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.817	0.970	0.893
3	1	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.812	0.972	0.892
4	2	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.811	0.967	0.889
5	3	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.787	0.950	0.868
6	3	Pterodactyl†	GPT-5.4	0.790	0.947	0.868
7	3	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.790	0.947	0.868
8	3	bytepop_pro9	Ensemble ^{E2}	0.787	0.948	0.867
9	3	bytepop_pro10	Ensemble ^{E2}	0.787	0.948	0.867
10	4	Vertical_2†	Gemma 3 1B IT	0.702	0.925	0.813
11	5	Vertical_1†	Gemma 3 1B IT	0.686	0.913	0.800
12	5	chrF	–	0.654	0.931	0.793
13	5	xCOMET XXL†	xCOMET-XXL	0.698	0.878	0.788
14	6	COMET22	XLM-RoBERTa-large	0.671	0.901	0.786
15	6	Cohere CAT+ ensemble*	Command A+	0.643	0.921	0.782
16	7	COMETKiwi22*	XLM-RoBERTa-large	0.687	0.853	0.770
17	7	xCOMET XXL*†	xCOMET-XXL	0.673	0.866	0.769
18	7	xCOMET XL†	xCOMET-XL	0.678	0.854	0.766
19	7	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.627	0.903	0.765
20	7	spBLEU	–	0.589	0.917	0.753
21	7	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.599	0.906	0.752
22	8	AEGIS	XLM-RoBERTa-large; mT5-large	0.642	0.842	0.742
23	8	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.545	0.924	0.734
24	9	xCOMET XL*†	xCOMET-XL	0.635	0.805	0.720
25	9	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.530	0.905	0.718
26	9	BERTScore	XLM-RoBERTa-large	0.533	0.885	0.709
27	9	MQM-LLM Confidence Aware*†	Not specified (27B)	0.519	0.893	0.706
28	9	gemba-poly*	Llama 3.3 70B	0.530	0.868	0.699
29	10	MQM-LLM*†	Not specified (27B)	0.459	0.912	0.686
30	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.450	0.919	0.684
31	10	BLEU	–	0.451	0.887	0.669
32	11	YiSi-1	XLM-RoBERTa-large	0.448	0.877	0.662
33	11	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.361	0.925	0.643
34	11	FACET*	GPT-4.1 Mini (2025-04-14)	0.361	0.925	0.643

Table 53: Task 2 ranking on the English→Thai translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.731	0.955	0.843
2	1	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.728	0.954	0.841
3	1	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.725	0.954	0.839
4	1	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.723	0.953	0.838
5	2	Pterodactyl*	GPT-5.4	0.720	0.943	0.831
6	2	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.720	0.943	0.831
7	2	bytepop_pro9	Ensemble ^{E2}	0.721	0.936	0.828
8	2	bytepop_pro10	Ensemble ^{E2}	0.721	0.936	0.828
9	2	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.706	0.950	0.828
10	3	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.688	0.939	0.814
11	3	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.682	0.945	0.813
12	4	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.657	0.951	0.804
13	5	Cohere CAT+ ensemble*	Command A+	0.645	0.917	0.781
14	5	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.624	0.936	0.780
15	5	MQM-LLM ^{*†}	Not specified (27B)	0.619	0.940	0.779
16	5	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.614	0.940	0.777
17	6	Vertical_1 [†]	Gemma 3 1B IT	0.609	0.921	0.765
18	6	Vertical_2 [†]	Gemma 3 1B IT	0.614	0.914	0.764
19	7	xCOMET XXL [†]	xCOMET-XXL	0.621	0.868	0.744
20	7	COMET22	XLM-RoBERTa-large	0.611	0.871	0.741
21	7	chrF	–	0.572	0.901	0.737
22	7	xCOMET XL [†]	xCOMET-XL	0.616	0.853	0.735
23	8	xCOMET XXL ^{*†}	xCOMET-XXL	0.589	0.845	0.717
24	8	spBLEU	–	0.526	0.895	0.711
25	8	AEGIS	XLM-RoBERTa-large; mT5-large	0.588	0.828	0.708
26	9	xCOMET XL ^{*†}	xCOMET-XL	0.565	0.822	0.693
27	9	BERTScore	XLM-RoBERTa-large	0.497	0.872	0.685
28	9	BLEU	–	0.486	0.881	0.684
29	9	COMETKiwi22*	XLM-RoBERTa-large	0.592	0.776	0.684
30	9	gemba-poly*	Llama 3.3 70B	0.480	0.884	0.682
31	10	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.401	0.908	0.655
32	11	YiSi-1	XLM-RoBERTa-large	0.397	0.868	0.633
33	11	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.364	0.900	0.632
34	11	FACET*	GPT-4.1 Mini (2025-04-14)	0.364	0.900	0.632

Table 54: Task 2 ranking on the English→Ukrainian translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.655	0.897	0.776
2	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.656	0.891	0.773
3	1	Gemma 4 - No Reasoning [*]	Gemma 4 31B IT	0.634	0.895	0.764
4	2	Gemma 4 - Reasoning [*]	Gemma 4 31B IT	0.640	0.888	0.764
5	2	Qwen 3.6 - Reasoning [*]	Qwen3.6-35B-A3B	0.627	0.895	0.761
6	2	Pterodactyl [†]	GPT-5.4	0.627	0.885	0.756
7	2	Austriadactylus [*]	GPT-5.4; GPT-5.6 Terra	0.627	0.885	0.756
8	2	bytepop_pro10	Ensemble ^{E2}	0.620	0.887	0.754
9	2	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.614	0.892	0.753
10	3	Gemini-3.6-Flash [*]	Gemini 3.6 Flash	0.602	0.873	0.738
11	3	bytepop_pro9	Ensemble ^{E2}	0.587	0.886	0.737
12	3	cuni-v14-regression [*]	Gemma 4 31B IT	0.566	0.901	0.734
13	3	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.554	0.905	0.730
14	4	Qwen 3.6 - No Reasoning [*]	Qwen3.6-35B-A3B	0.490	0.907	0.699
15	4	Cohere CAT+ ensemble [†]	Command A+	0.521	0.855	0.688
16	4	Vertical_2 [†]	Gemma 3 1B IT	0.473	0.885	0.679
17	4	fluency2-gemini35-esa [*]	Gemini 3.5 Flash	0.460	0.883	0.671
18	4	xCOMET XXL [†]	xCOMET-XXL	0.488	0.848	0.668
19	5	Vertical_1 [†]	Gemma 3 1B IT	0.444	0.862	0.653
20	5	xCOMET XXL ^{*†}	xCOMET-XXL	0.456	0.839	0.647
21	5	xCOMET XL ^{*†}	xCOMET-XL	0.427	0.805	0.616
22	5	FACET [*]	GPT-4.1 Mini (2025-04-14)	0.353	0.876	0.615
23	5	doc-FACET [*]	GPT-4.1 Mini (2025-04-14)	0.353	0.876	0.615
24	5	gemba-poly [*]	Llama 3.3 70B	0.410	0.812	0.611
25	5	xCOMET XL [†]	xCOMET-XL	0.425	0.795	0.610
26	6	COMETKiwi22 [*]	XLM-RoBERTa-large	0.442	0.713	0.578
27	6	AEGIS	XLM-RoBERTa-large; mT5-large	0.385	0.735	0.560
28	6	COMET22	XLM-RoBERTa-large	0.365	0.749	0.557
29	6	MQM-LLM ^{*†}	Not specified (27B)	0.262	0.811	0.536
30	6	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.266	0.805	0.535
31	7	YiSi-1	BERT-base-chinese	0.300	0.697	0.499
32	8	BERTScore	XLM-RoBERTa-large	0.276	0.686	0.481
33	8	chrF	—	0.230	0.701	0.465
34	8	BLEU	—	0.211	0.705	0.458
35	8	spBLEU	—	0.204	0.709	0.457

Table 55: Task 2 ranking on the English→Chinese (Simplified) translation direction. ^{*} marks reference-free submissions. [†] marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.605	0.849	0.727
2	1	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.598	0.852	0.725
3	1	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.586	0.854	0.720
4	2	bytopop_pro10	Ensemble ^{E2}	0.576	0.836	0.706
5	2	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.559	0.852	0.706
6	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.564	0.848	0.706
7	2	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.554	0.854	0.704
8	2	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.552	0.846	0.699
9	2	Pterodactyl*	GPT-5.4	0.551	0.843	0.697
10	2	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.551	0.843	0.697
11	3	bytopop_pro9	Ensemble ^{E2}	0.558	0.829	0.693
12	3	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.525	0.860	0.693
13	3	cuni-v14-regression*	Gemma 4 31B IT	0.518	0.834	0.676
14	4	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.504	0.822	0.663
15	4	Cohere CAT+ ensemble*	Command A+	0.502	0.820	0.661
16	5	xCOMET XXL [†]	xCOMET-XXL	0.444	0.780	0.612
17	5	Vertical_2 [†]	Gemma 3 1B IT	0.407	0.802	0.604
18	5	BLEU	–	0.389	0.795	0.592
19	5	gemba-poly*	Llama 3.3 70B	0.357	0.825	0.591
20	5	xCOMET XXL ^{*†}	xCOMET-XXL	0.389	0.787	0.588
21	6	Vertical_1 [†]	Gemma 3 1B IT	0.380	0.794	0.587
22	6	xCOMET XL [†]	xCOMET-XL	0.436	0.734	0.585
23	6	chrF	–	0.389	0.778	0.583
24	6	spBLEU	–	0.384	0.781	0.582
25	6	COMET22	XLM-RoBERTa-large	0.404	0.753	0.578
26	6	COMETKiwi22*	XLM-RoBERTa-large	0.430	0.726	0.578
27	6	xCOMET XL ^{*†}	xCOMET-XL	0.394	0.751	0.572
28	6	YiSi-1	BERT-base-chinese	0.362	0.770	0.566
29	6	FACET*	GPT-4.1 Mini (2025-04-14)	0.311	0.817	0.564
30	6	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.311	0.817	0.564
31	6	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.316	0.811	0.563
32	6	BERTScore	XLM-RoBERTa-large	0.348	0.774	0.561
33	6	AEGIS	XLM-RoBERTa-large; mT5-large	0.390	0.728	0.559
34	7	MQM-LLM ^{*†}	Not specified (27B)	0.243	0.749	0.496
35	7	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.231	0.734	0.483

Table 56: Task 2 ranking on the English→Chinese (Traditional, Taiwan) translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Rank	Cluster	Auto-evaluator	Underlying model	PDP	SPA	Avg Corr
1	1	Gemma 4 Reference-based - Reasoning	Gemma 4 31B IT	0.771	0.940	0.856
2	1	GPT-5.6 Sol - Medium Reasoning*	GPT-5.6 Sol	0.757	0.947	0.852
3	2	Gemma 4 - Reasoning*	Gemma 4 31B IT	0.758	0.940	0.849
4	2	Gemma 4 Reference-based - No Reasoning	Gemma 4 31B IT	0.763	0.933	0.848
5	2	Gemini-3.6-Flash*	Gemini 3.6 Flash	0.756	0.939	0.848
6	2	GPT-5.6 Terra*	GPT-5.6 Terra	0.748	0.942	0.845
7	3	Gemma 4 - No Reasoning*	Gemma 4 31B IT	0.737	0.935	0.836
8	3	cuni-da-6d*	Gemma 4 31B IT	0.732	0.925	0.828
9	4	Pterodactyl*	GPT-5.4	0.711	0.924	0.817
10	4	Austriadactylus*	GPT-5.4; GPT-5.6 Terra	0.711	0.924	0.817
11	4	Qwen 3.6 Reference-based - Reasoning	Qwen3.6-35B-A3B	0.700	0.931	0.815
12	4	bytepop_pro9	Ensemble ^{E2}	0.701	0.929	0.815
13	4	bytepop_pro10	Ensemble ^{E2}	0.701	0.929	0.815
14	4	cuni-v14-regression*	Gemma 4 31B IT	0.689	0.931	0.810
15	4	Vertical_2 [†]	Gemma 3 1B IT	0.693	0.917	0.805
16	5	Qwen 3.6 - Reasoning*	Qwen3.6-35B-A3B	0.668	0.930	0.799
17	5	Vertical_1 [†]	Gemma 3 1B IT	0.674	0.901	0.788
18	6	chrF	–	0.644	0.912	0.778
19	6	Qwen 3.6 Reference-based - No Reasoning	Qwen3.6-35B-A3B	0.601	0.937	0.769
20	7	COMET22	XLM-RoBERTa-large	0.634	0.881	0.758
21	7	BLEU	–	0.600	0.912	0.756
22	7	MQM-LLM ^{*†}	Not specified (27B)	0.585	0.922	0.754
23	8	MQM-LLM Confidence Aware ^{*†}	Not specified (27B)	0.576	0.918	0.747
24	8	spBLEU	–	0.576	0.915	0.746
25	8	Qwen 3.6 - No Reasoning*	Qwen3.6-35B-A3B	0.537	0.919	0.728
26	9	AEGIS	XLM-RoBERTa-large; mT5-large	0.603	0.848	0.725
27	9	BERTScore	XLM-RoBERTa-large	0.547	0.892	0.720
28	9	Lexicala-QE-Ensemble (Majority-Vote)*	Ensemble ^{E3}	0.524	0.908	0.716
29	10	Cohere CAT+ ensemble*	Command A+	0.482	0.925	0.703
30	10	gemba-poly*	Llama 3.3 70B	0.474	0.915	0.694
31	10	xCOMET XXL [†]	xCOMET-XXL	0.513	0.871	0.692
32	11	xCOMET XL [†]	xCOMET-XL	0.496	0.866	0.681
33	12	Lexicala-QE-Ensemble (STAPLE)*	Ensemble ^{E3}	0.418	0.902	0.660
34	12	YiSi-1	XLM-RoBERTa-large	0.410	0.875	0.642
35	13	PhraseT2-01*	Gemini 3.5 Flash; Gemma 4 12B	0.334	0.920	0.627
36	13	fluency2-gemini35-esa*	Gemini 3.5 Flash	0.340	0.907	0.624
37	14	COMETKiwi22*	XLM-RoBERTa-large	0.422	0.766	0.594
38	14	PhraseT2-02*	Ensemble ^{E1}	0.264	0.907	0.585
39	14	xCOMET XXL ^{*†}	xCOMET-XXL	0.357	0.805	0.581
40	15	FACET*	GPT-4.1 Mini (2025-04-14)	0.233	0.874	0.554
41	15	doc-FACET*	GPT-4.1 Mini (2025-04-14)	0.233	0.874	0.554
42	15	xCOMET XL ^{*†}	xCOMET-XL	0.317	0.741	0.529

Table 57: Task 2 ranking on the Chinese (Simplified)→Japanese translation direction. * marks reference-free submissions. † marks fine-tuned systems. Light-grey rows are organizer baselines. Ensemble components named in the model cards are: E1 [Gemini 3.5 Flash; Gemma 4 12B; Gemma 4 31B]; E2 [GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; COMET-22]; E3 [Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus]. Statistical significance clusters use two-sided source-level permutation tests (1,000 draws per evaluator pair, unadjusted $p \leq 0.05$), following the resampling framework of Deutsch et al. (2021b); cluster labels use the greedy procedure adopted in recent WMT Metrics evaluations (Freitag et al., 2024; Lavie et al., 2025).

Direction	PDP ρ	SPA ρ
Czech→German	0.989	0.838
Czech→Vietnamese	0.993	0.884
English→Arabic (Egyptian)	0.994	0.946
English→Estonian	0.995	0.731
English→Indonesian	0.995	0.839
English→Icelandic	0.994	0.744
English→Northern Sámi	0.992	0.959
Chinese→Japanese	0.998	0.886

Table 58: Spearman correlations between auto-evaluator rankings under screen- and source-based grouping. PDP compares screen-based with source-based PDP; SPA compares screen-based with source-based SPA. Each row uses only direct Task 2 submissions and direct Task 2 baselines; Task 1-to-Task 2 conversions are excluded.

E Baseline Prompting Details

The LLM-based baselines follow the two-stage GEMBA-ESA prompting strategy. In Stage 1 (Task 1), the model identifies and classifies translation errors using MQM-style categories. In Stage 2 (Task 2), the model assigns a quality score on a 0–100 scale given the source, translation, and the error annotations produced by Stage 1. Both stages use a shared system prompt:

Your task is to identify machine translation errors and assess the quality of the translation.

Each stage is issued as a single-turn chat message (system + user). Stage 1 uses chain-of-thought decoding when enabled (temperature 0.6, top- p 0.95, max 8192 tokens); otherwise greedy decoding (temperature 0). Stage 2 always uses greedy decoding with a 64-token limit and thinking disabled.

E.1 Stage 1: Error Annotation (Task 1)

Figure 7 shows the Stage 1 prompt template. When a reference translation is available (`-with-ref`), the shaded reference block is inserted between the translation and the annotation instructions; otherwise it is omitted. The reference header adapts to the reference type (Table 64).

E.2 Stage 2: Quality Scoring (Task 2)

Figure 8 shows the Stage 2 prompt template. The `{error_spans}` placeholder is filled with the parsed Stage 1 output. As with Stage 1, the reference block is included only when `-with-ref` is active.

E.3 Domain Requirements

Each segment carries a domain label extracted from its `item_id`. The domain-specific instruction text appended to both Stage 1 and Stage 2 prompts is shown in Table 63.

E.4 Reference Type Labels

When the reference-augmented variant is used, the reference block header adapts to the type of reference available for each segment, as shown in Table 64.

F Task 3

F.1 Details on Statistical Significance Tests

To determine the statistical significance for a certain pair of systems, A and B, the observed test

statistic is the difference between their macro-average MCCs. Each macro-average is the unweighted mean of eight MCCs computed separately for the eight language pairs; it is not an MCC computed after pooling their observations.

Prior permutation procedures used in MT metric evaluation (Deutsch et al., 2021b; Thompson et al., 2024) do not fit our macro-average setting exactly, so we adapt them. Our procedure is most similar to PERM-INPUTS from Deutsch et al. (2021b), which makes one exchange decision per input and applies it to all system outputs for that input. For each pair of metrics, we jointly exchange their predictions at the source-segment level, using the same decision for every observation and language-pair occurrence associated with that source segment.

If a segment is retained, A keeps all of A’s predictions for that segment and B keeps all of B’s. If it is exchanged, A receives all of B’s predictions for that segment and B receives all of A’s. Each decision therefore covers every MT-system output and repeated human judgment associated with the segment. When the same source segment occurs in multiple language pairs, the same exchange decision is used for every occurrence. The 2,357 language-pair-specific items form 1,392 source-segment units, of which 553 occur in more than one priority pair. Units are identified by set, source language, domain, document and segment; target language is excluded, and shared units were verified to have identical source text.

After each exchange, MCC is recomputed separately for each language pair, and the eight resulting MCCs are averaged. For each of the 153 unordered metric pairs, 100,000 permutations were drawn.

To convert the pairwise permutation-test results into clusters, we adopt the greedy clustering procedure used in recent WMT evaluation shared tasks (Freitag et al., 2022, 2023, 2024; Lavie et al., 2025). We use the unadjusted two-sided Monte Carlo p -value and the clustering threshold is $p < 0.05$. The highest metric starts cluster 1. Moving down the ranking, a metric joins the current cluster only when its pairwise difference from every metric already in that cluster is not significant. If it differs significantly from at least one current-cluster metric, it starts the next numbered cluster. Thus every pair within a cluster is nonsignificant under these tests, while membership in different clusters does not imply that every cross-cluster pair differs significantly.

Stage 1 Prompt Template (User Turn)

```

{src_name} source:
```{src_text}```
{tgt_name} translation:
```{hyp_text}```

```

Inserted only when `-with-ref` is active

```

{ref_label}:
```{ref_text}```

```

Based on the source segment{ , machine translation, and reference translation | and machine translation} surrounded by triple backticks, identify error types in the translation and classify them. The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error.

Each error is classified as one of two categories: major or minor. Major errors disrupt the flow and make the understandability of the text difficult or impossible. Minor errors are errors that do not disrupt the flow significantly, and what the text is trying to say is still understandable.

Return only the annotations in this format:

Major:  
category/subcategory - “error span”

Minor:  
category/subcategory - “error span”

Use one error per line and write no-error when a section is empty. Quote spans from the translation; for omissions, quote the omitted source span.

Domain requirements: {domain\_req}

Figure 7: Stage 1 (Task 1) error annotation prompt template. Placeholders in braces are filled per-segment. The gray inner box is included only in the reference-augmented variant. The instruction wording adapts accordingly (gray alternation text).

Stage 2 Prompt Template (User Turn)

Given the translation from {src\_name} to {tgt\_name} and the annotated error spans, assign a score on a continuous scale from 0 to 100. The scale has the following reference points: 0=“No meaning preserved”, 33=“Some meaning preserved”, 66=“Most meaning preserved and few grammar mistakes”, up to 100=“Perfect meaning and grammar”.

Domain requirements: {domain\_req}

Score the following translation:

```

{src_name} source:
```{src_text}```
{tgt_name} translation:
```{hyp_text}```

```

Inserted only when `-with-ref` is active

```

{ref_label}:
```{ref_text}```

```

Annotated error spans:

```

```{error_spans}```

```

Respond with ONLY a valid JSON object and nothing else: {"score": N}  
where N is an integer from 0 to 100.

Figure 8: Stage 2 (Task 2) quality scoring prompt template. The {error\_spans} field contains the Stage 1 annotation output. The gray inner box is included only in the reference-augmented variant.

Table 63: Domain requirement text inserted into both Stage 1 and Stage 2 prompts. The *news* and *factchecking* domains share identical requirements.

Domain	Requirement text
news / factchecking	The source segment is from a news article. The translation should use a formal register consistent with journalistic standards and preserve the source HTML formatting.
speech	The source segment is a transcript of spoken content from a video. The translation should preserve the speaker’s flow and colloquial style. It should omit non-linguistic sounds, such as laughter, groans, and hesitation sounds, while retaining interjections. Interrupted words should be completed when they can be inferred from context; otherwise, they should be omitted. Foreign words should remain unchanged. Each sentence should be placed on a separate line.
social	The source segment is user-generated content from a social media platform. Source spelling mistakes should not be reproduced. Meaningful expressiveness, such as capitalization or elongation, should be reproduced naturally in the target language. URLs and user handles should be copied unchanged, while hashtags should be translated when appropriate. Source punctuation should be followed as closely as possible, with additional punctuation only when needed to prevent serious loss of comprehension. The translation should use an informal style, like close friends talking, even if this changes the original tone, and preserve the source HTML formatting.
software	The source segment contains software data from a JSON. Only JSON content or values should be translated; keys and placeholders should be copied unchanged. The translation should contain only valid JSON content matching the input format.
edu	The source segment consists of biology, chemistry, and geography exercises from an educational web portal for children aged 9–16. The translation should be suitable for this educational context and age range, and preserve the source HTML formatting.
general	The translation should be accurate and fluent.

Table 64: Reference type labels used in prompt headers.

Type	Prompt label
human	Human reference translation
postedit	Post-edited reference translation
pseudo	High-quality machine translation reference

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Austriadactylus	Ensemble <sup>E1</sup>	T2	24.27	38.75	80.73	52.36
1	Pterodactyl	GPT-5.4	T2	24.27	38.75	80.73	52.36
3	Wukongopterus	GPT-5.4	T3	23.92	37.39	86.63	52.24
4	PhraseT3-02	Gemini 3.5 Flash	T3	23.71	38.68	79.67	52.08
5	Austriadraco	GPT-5.4	T3	23.60	38.95	78.05	51.97
6	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	22.82	46.45	45.16	45.79
7	PhraseT3-01	Gemini 3.5 Flash	T3	22.68	37.26	84.51	51.71
8	cuni-cat-v4	Gemma 4 31B	T1	22.57	40.38	68.01	50.68
9	cuni-v14	Gemma 4 31B	T1	22.47	40.15	68.87	50.72
10	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T2	22.16	39.50	71.24	50.82
11	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	22.00	43.22	53.73	47.91
12	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	21.97	46.57	42.43	44.40
13	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	21.79	41.91	58.53	48.84
14	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	21.70	42.51	55.70	48.22
15	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	21.63	44.08	49.39	46.59
16	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E3</sup>	T2	21.00	40.13	64.48	49.47
17	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T1	20.86	41.43	57.77	48.25
18	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	20.33	42.72	50.50	46.29
19	AEGIS	Ensemble <sup>E4</sup>	T2	19.55	45.38	38.90	41.89
20	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	19.54	40.72	56.71	47.41
21	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T3	19.45	42.60	48.18	45.22
22	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	18.93	44.48	39.81	42.01
23	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E3</sup>	T1	18.28	44.95	36.38	40.21
24	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E3</sup>	T3	17.95	46.03	32.49	38.10
25	RAGthoven Error-Free	Ensemble <sup>E5</sup>	T3	16.55	35.37	80.02	49.06
26	gemba-poly	Llama 3.3 70B	T2	16.35	40.66	45.71	43.04
27	Vertical_2	Gemma 3 1B IT	T2	16.16	36.49	70.28	48.03
28	Cohere CAT+ ensemble	Command A+	T2	15.18	49.80	18.52	27.00
29	Cohere CAT+	Command A+	T1	14.75	44.96	25.88	32.85
30	Vertical_1	Gemma 3 1B IT	T2	12.81	36.70	54.69	43.92
31	First try [871721]	Qwen3.6-27B	T3	12.67	41.46	28.91	34.07
32	First try [874374]	Babel-83B-Chat	T3	10.89	31.73	96.01	47.70
33	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	10.66	45.76	13.62	21.00
34	bytepop_pro9	Ensemble <sup>E6</sup>	T2	9.13	30.88	99.60	47.15
34	bytepop_pro10	Ensemble <sup>E7</sup>	T2	9.13	30.88	99.60	47.15
36	FACET	GPT-4.1 Mini (2025-04-14)	T1	7.52	40.17	14.23	21.01
37	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	7.49	40.17	14.13	20.90
38	FACET	GPT-4.1 Mini (2025-04-14)	T2	7.24	40.06	13.52	20.22
38	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	7.24	40.06	13.52	20.22
40	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	7.14	40.00	13.32	19.98
40	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	7.14	40.00	13.32	19.98
42	Qwen-based Omission	Qwen3-32B	T1	2.95	32.05	36.93	34.32
43	Rule-based Omission	Qwen3-32B	T1	2.94	32.04	37.03	34.36
44	xCOMET XL	xCOMET-XL	T1-B	-1.40	27.85	6.41	10.42
45	xCOMET XXL	xCOMET-XXL	T1-B	-7.66	23.05	14.88	18.09

Table 65: WMT26 Task 3 results for ces-deu (ces\_Latn--deu\_Latn). 438/438 items (100.0%) and 5,909/13,578 item-system pairs (43.5%) are annotated, giving 6,564 human assessments (1,982 error-free, 4,582 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> Lexicala-QE-Ensemble (Majority-Vote): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E3</sup> Lexicala-QE-Ensemble (STAPLE): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E4</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E5</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro; <sup>E6</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E7</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Wukongopterus	GPT-5.4	T3	29.19	43.71	84.91	57.71
2	Austriadactylus	Ensemble <sup>E1</sup>	T2	27.82	44.89	77.18	56.76
2	Pterodactyl	GPT-5.4	T2	27.82	44.89	77.18	56.76
4	Austriadraco	GPT-5.4	T3	27.75	45.42	74.85	56.54
5	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	26.66	52.54	48.13	50.24
6	cuni-cat-v4	Gemma 4 31B	T1	25.57	46.19	66.20	54.41
7	cuni-v14	Gemma 4 31B	T1	25.51	46.14	66.25	54.40
8	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	24.97	49.00	54.19	51.46
9	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	24.72	52.82	42.56	47.14
10	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	24.39	48.42	54.57	51.31
11	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	22.35	51.31	40.24	45.10
12	First try [871721]	Qwen3.6-27B	T3	20.59	50.03	38.83	43.73
13	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	19.50	50.31	35.10	41.35
14	bytepop_pro9	Ensemble <sup>E2</sup>	T2	18.77	36.23	99.35	53.10
14	bytepop_pro10	Ensemble <sup>E3</sup>	T2	18.77	36.23	99.35	53.10
16	gemba-poly	Llama 3.3 70B	T2	18.62	45.00	50.35	47.52
17	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	18.55	49.10	35.59	41.27
18	Cohere CAT+ ensemble	Command A+	T2	18.54	59.08	18.66	28.36
19	First try [874374]	Babel-83B-Chat	T3	16.09	36.51	95.19	52.78
20	AEGIS	Ensemble <sup>E4</sup>	T2	16.05	53.59	20.17	29.31
21	Vertical_2	Gemma 3 1B IT	T2	15.64	39.04	75.99	51.58
22	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	15.12	59.84	12.33	20.45
23	Vertical_1	Gemma 3 1B IT	T2	14.01	38.12	78.04	51.22
24	Cohere CAT+	Command A+	T1	13.44	46.76	26.50	33.83
25	xCOMET XXL	xCOMET-XXL	T1-B	6.51	39.65	24.66	30.41
26	xCOMET XL	xCOMET-XL	T1-B	6.01	41.24	15.14	22.15
27	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	1.58	38.06	3.19	5.89
28	FACET	GPT-4.1 Mini (2025-04-14)	T1	1.50	37.82	3.19	5.89
29	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	0.30	34.44	3.35	6.11
29	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	0.30	34.44	3.35	6.11
31	FACET	GPT-4.1 Mini (2025-04-14)	T2	0.15	34.07	3.35	6.11
31	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	0.15	34.07	3.35	6.11

Table 66: WMT26 Task 3 results for ces-ukr (ces\_Latn--ukr\_Cyrl). 415/415 items (100.0%) and 4,976/12,035 item–system pairs (41.3%) are annotated, giving 5,490 human assessments (1,849 error-free, 3,641 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E3</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> AEGIS: XLM-RoBERTa-large; mT5-large.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	28.10	42.51	56.75	48.61
2	Wukongopterus	GPT-5.4	T3	27.71	36.04	80.00	49.69
3	cuni-cat-v4	Gemma 4 31B	T1	25.79	37.85	67.00	48.38
4	cuni-v14	Gemma 4 31B	T1	25.54	37.62	67.25	48.25
5	Austriadactylus	Ensemble <sup>E1</sup>	T2	25.29	37.40	67.50	48.13
5	Pterodactyl	GPT-5.4	T2	25.29	37.40	67.50	48.13
7	Austriadraco	GPT-5.4	T3	24.87	37.76	64.75	47.70
8	PhraseT3-01	Gemini 3.5 Flash	T3	24.28	34.90	76.25	47.88
9	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	24.24	42.96	45.00	43.96
10	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	24.15	46.75	36.00	40.68
11	PhraseT3-02	Gemini 3.5 Flash	T3	23.20	35.17	71.75	47.20
12	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	23.06	43.98	39.25	41.48
13	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	22.48	43.54	38.75	41.01
14	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	22.35	42.90	40.00	41.40
15	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	21.55	44.82	33.50	38.34
16	gemba-poly	Llama 3.3 70B	T2	21.33	37.90	53.25	44.28
17	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	18.11	44.84	25.00	32.10
18	First try [874374]	Babel-83B-Chat	T3	16.55	28.69	93.75	43.94
19	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	14.61	37.54	32.00	34.55
20	Cohere CAT+	Command A+	T1	13.49	38.89	24.50	30.06
21	First try [871721]	Qwen3.6-27B	T3	13.39	41.67	18.75	25.86
22	Cohere CAT+ ensemble	Command A+	T2	11.37	53.06	6.50	11.58
23	AEGIS	Ensemble <sup>E2</sup>	T2	11.07	42.37	12.50	19.31
24	bytepop_pro9	Ensemble <sup>E3</sup>	T2	8.24	26.21	98.50	41.41
24	bytepop_pro10	Ensemble <sup>E4</sup>	T2	8.24	26.21	98.50	41.41
26	RAGthoven Error-Free	Ensemble <sup>E5</sup>	T3	7.41	30.41	35.50	32.76
27	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	5.48	35.90	7.00	11.72
28	Vertical_2	Gemma 3 1B IT	T2	4.84	36.21	5.25	9.17
29	Vertical_1	Gemma 3 1B IT	T2	4.32	30.81	13.25	18.53
30	Qwen-based Omission	Qwen3-32B	T1	2.83	27.25	34.00	30.26
31	Rule-based Omission	Qwen3-32B	T1	2.59	27.09	34.00	30.16
32	FACET	GPT-4.1 Mini (2025-04-14)	T1	1.12	28.26	3.25	5.83
32	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	1.12	28.26	3.25	5.83
34	FACET	GPT-4.1 Mini (2025-04-14)	T2	-0.39	24.53	3.25	5.74
34	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	-0.39	24.53	3.25	5.74
34	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	-0.39	24.53	3.25	5.74
34	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	-0.39	24.53	3.25	5.74
38	XCOMOffset	—	T1	-0.83	24.80	23.25	24.00
39	xCOMET XL	xCOMET-XL	T1-B	-6.53	19.44	14.00	16.28
40	xCOMET XXL	xCOMET-XXL	T1-B	-6.81	22.25	40.50	28.72

Table 67: WMT26 Task 3 results for ces-vie (ces\_Latn--vie\_Latn). 49/418 items (11.7%) and 1,378/12,958 item-system pairs (10.6%) are annotated, giving 1,572 human assessments (400 error-free, 1,172 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E5</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	bytepop_pro9	Ensemble <sup>E1</sup>	T2	36.24	66.54	98.46	79.41
1	bytepop_pro10	Ensemble <sup>E2</sup>	T2	36.24	66.54	98.46	79.41
3	Wukongopterus	GPT-5.4	T3	32.21	75.34	65.82	70.26
4	Austriadactylus	Ensemble <sup>E3</sup>	T2	28.24	76.94	52.46	62.39
4	Pterodactyl	GPT-5.4	T2	28.24	76.94	52.46	62.39
6	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E4</sup>	T2	28.09	77.47	50.53	61.16
7	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E4</sup>	T1	27.65	78.81	45.84	57.96
8	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E5</sup>	T2	27.32	78.42	46.18	58.13
9	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	27.24	81.88	37.75	51.67
9	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	27.24	81.88	37.75	51.67
11	Austriadraco	GPT-5.4	T3	26.75	77.36	47.89	59.16
12	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E5</sup>	T1	26.03	80.63	37.97	51.63
13	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	25.47	77.31	45.16	57.01
14	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	24.38	79.80	36.34	49.94
15	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	23.53	79.39	35.40	48.97
16	cuni-cat-v4	Gemma 4 31B	T1	23.44	75.99	44.30	55.97
17	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	22.83	79.57	33.46	47.11
18	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E4</sup>	T3	22.22	81.27	28.78	42.51
19	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E5</sup>	T3	22.12	82.40	26.64	40.26
20	cuni-v14	Gemma 4 31B	T1	22.04	74.84	44.52	55.82
21	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	21.36	80.35	28.71	42.31
22	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	20.77	80.33	27.50	40.98
23	AEGIS	Ensemble <sup>E6</sup>	T2	20.21	84.50	20.18	32.58
24	PhraseT3-01	Gemini 3.5 Flash	T3	19.93	71.13	53.45	61.04
25	PhraseT3-02	Gemini 3.5 Flash	T3	19.44	69.94	57.95	63.38
26	Cohere CAT+	Command A+	T1	18.73	76.26	31.70	44.79
27	gemba-poly	Llama 3.3 70B	T2	18.12	80.43	21.87	34.39
28	Cohere CAT+ ensemble	Command A+	T2	17.04	85.67	13.95	23.99
29	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	16.20	77.17	23.40	35.91
30	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	14.92	81.49	14.48	24.59
31	FACET	GPT-4.1 Mini (2025-04-14)	T1	14.91	81.41	14.55	24.68
32	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	14.66	79.07	16.88	27.82
33	First try [874374]	Babel-83B-Chat	T3	13.28	67.82	49.03	56.91
34	First try [871721]	Qwen3.6-27B	T3	13.28	79.26	13.97	23.75
35	RAGthoven Error-Free	Ensemble <sup>E7</sup>	T3	12.96	69.21	39.34	50.16
36	XCOMOffset	—	T1	12.56	68.22	43.42	53.06
37	xCOMET XXL	xCOMET-XXL	T1-B	11.29	71.28	24.35	36.30
38	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	9.62	73.25	14.42	24.10
38	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	9.62	73.25	14.42	24.10
40	FACET	GPT-4.1 Mini (2025-04-14)	T2	9.39	72.86	14.50	24.19
40	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	9.39	72.86	14.50	24.19
42	Vertical_2	Gemma 3 1B IT	T2	5.95	71.14	8.13	14.60
43	Vertical_1	Gemma 3 1B IT	T2	5.52	68.36	11.79	20.11
44	Qwen-based Omission	Qwen3-32B	T1	1.00	61.22	29.04	39.39
45	Rule-based Omission	Qwen3-32B	T1	0.85	61.10	29.14	39.46
46	xCOMET XL	xCOMET-XL	T1-B	-13.85	37.52	4.98	8.79

Table 68: WMT26 Task 3 results for eng-arz (eng\_Latn--arz\_Arab). 433/433 items (100.0%) and 12,990/12,990 item-system pairs (100.0%) are annotated, giving 14,886 human assessments (8,999 error-free, 5,887 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E2</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E3</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E4</sup> Lexicala-QE-Ensemble (Majority-Vote): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E5</sup> Lexicala-QE-Ensemble (STAPLE): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E6</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E7</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	45.93	48.02	73.44	58.07
2	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	44.71	48.86	69.01	57.21
3	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	44.59	48.06	70.41	57.13
4	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	44.24	52.51	61.07	56.47
5	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	41.92	49.83	61.00	54.85
6	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	41.89	40.32	83.27	54.33
7	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	41.16	45.40	68.59	54.63
8	Austriadraco	GPT-5.4	T3	40.57	38.23	86.30	52.99
9	Cohere CAT+ ensemble	Command A+	T2	39.75	54.28	48.98	51.50
10	Austriadactylus	Ensemble <sup>E1</sup>	T2	39.70	37.10	87.77	52.15
10	Pterodactyl	GPT-5.4	T2	39.70	37.10	87.77	52.15
12	Wukongopterus	GPT-5.4	T3	37.82	34.26	92.69	50.03
13	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	37.81	55.96	42.87	48.55
14	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	36.61	56.59	39.85	46.76
15	FACET	GPT-4.1 Mini (2025-04-14)	T1	36.58	56.53	39.85	46.74
16	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	34.70	53.19	40.48	45.97
16	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	34.70	53.19	40.48	45.97
18	FACET	GPT-4.1 Mini (2025-04-14)	T2	34.57	52.99	40.48	45.90
18	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	34.57	52.99	40.48	45.90
20	Cohere CAT+	Command A+	T1	33.88	43.11	56.71	48.98
21	First try [871721]	Qwen3.6-27B	T3	31.92	50.41	38.72	43.80
22	gemba-poly	Llama 3.3 70B	T2	23.57	38.07	42.59	40.20
23	xCOMET XXL	xCOMET-XXL	T1-B	21.22	40.36	31.20	35.20
24	AEGIS	Ensemble <sup>E2</sup>	T2	19.36	42.49	23.47	30.24
25	Vertical_2	Gemma 3 1B IT	T2	18.92	32.54	46.24	38.20
26	First try [874374]	Babel-83B-Chat	T3	18.23	25.42	86.93	39.33
27	bytepop_pro10	Ensemble <sup>E3</sup>	T2	16.78	22.89	100.00	37.26
28	bytepop_pro9	Ensemble <sup>E4</sup>	T2	15.86	22.66	100.00	36.94
29	xCOMET XL	xCOMET-XL	T1-B	15.64	41.80	16.65	23.82
30	Vertical_1	Gemma 3 1B IT	T2	15.07	33.54	29.73	31.52
31	PTT-MULTITAN-v2 [875519]	—	T3	14.11	23.79	88.55	37.51
32	PTT-MULTITAN-v2 [875520]	—	T3	8.87	25.10	48.00	32.96

Table 69: WMT26 Task 3 results for eng-bel (eng\_Latn--bel\_Cyrl). 354/354 items (100.0%) and 6,153/10,974 item–system pairs (56.1%) are annotated, giving 6,888 human assessments (1,423 error-free, 5,465 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	40.84	56.30	60.01	58.10
2	Cohere CAT+ ensemble	Command A+	T2	39.65	65.51	43.36	52.18
3	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	39.36	50.05	70.64	58.59
4	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	39.12	58.95	51.76	55.12
5	Cohere CAT+	Command A+	T1	38.50	57.27	53.41	55.27
6	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	38.12	59.96	48.17	53.42
7	cuni-cat-v4	Gemma 4 31B	T1	37.87	47.84	73.15	57.85
8	cuni-v14	Gemma 4 31B	T1	37.66	47.46	73.73	57.75
9	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	36.59	53.24	57.21	55.16
10	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	36.11	51.72	59.51	55.34
11	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	35.35	51.95	57.29	54.49
12	Austriadraco	GPT-5.4	T3	35.35	44.47	77.39	56.48
13	Austriadactylus	Ensemble <sup>E1</sup>	T2	35.13	43.42	80.26	56.35
13	Pterodactyl	GPT-5.4	T2	35.13	43.42	80.26	56.35
15	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	32.95	63.37	34.03	44.28
16	Wukongopterus	GPT-5.4	T3	32.80	40.05	86.72	54.80
17	First try [871721]	Qwen3.6-27B	T3	32.76	58.27	40.20	47.58
18	xCOMET XXL	xCOMET-XXL	T1-B	32.58	54.23	46.52	50.08
19	AEGIS	Ensemble <sup>E2</sup>	T2	30.12	60.82	31.87	41.83
20	FACET	GPT-4.1 Mini (2025-04-14)	T1	25.28	57.29	27.06	36.76
21	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	25.20	57.23	26.99	36.68
22	gemba-poly	Llama 3.3 70B	T2	24.74	47.11	43.29	45.12
23	FACET	GPT-4.1 Mini (2025-04-14)	T2	22.71	53.71	26.49	35.48
23	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	22.71	53.71	26.49	35.48
23	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	22.71	53.71	26.49	35.48
23	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	22.71	53.71	26.49	35.48
27	First try [874374]	Babel-83B-Chat	T3	17.50	32.87	87.01	47.72
28	bytepop_pro9	Ensemble <sup>E3</sup>	T2	16.17	30.21	99.93	46.39
28	bytepop_pro10	Ensemble <sup>E4</sup>	T2	16.17	30.21	99.93	46.39
30	xCOMET XL	xCOMET-XL	T1-B	15.36	53.09	13.57	21.61
31	Vertical_2	Gemma 3 1B IT	T2	9.34	44.63	9.83	16.12
32	Vertical_1	Gemma 3 1B IT	T2	7.69	37.97	15.29	21.80

Table 70: WMT26 Task 3 results for eng-ces (eng\_Latn--ces\_Latn). 326/448 items (72.8%) and 4,490/13,440 item-system pairs (33.4%) are annotated, giving 4,920 human assessments (1,393 error-free, 3,527 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	40.58	57.60	63.66	60.48
2	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	39.96	61.94	54.10	57.76
3	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	39.18	58.21	59.56	58.88
4	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	38.83	52.11	72.97	60.80
5	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	38.64	56.73	61.55	59.04
6	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	38.33	60.04	54.27	57.01
7	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	37.65	68.26	40.57	50.89
8	AEGIS	Ensemble <sup>E1</sup>	T2	37.27	63.33	46.62	53.70
9	Austriadraco	GPT-5.4	T3	35.89	46.34	83.59	59.63
10	Cohere CAT+ ensemble	Command A+	T2	35.80	69.16	36.34	47.64
11	Austriadactylus	Ensemble <sup>E2</sup>	T2	35.50	45.40	85.79	59.38
11	Pterodactyl	GPT-5.4	T2	35.50	45.40	85.79	59.38
13	Wukongopterus	GPT-5.4	T3	34.73	42.95	92.17	58.60
14	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	34.65	54.49	58.04	56.21
15	Cohere CAT+	Command A+	T1	33.90	63.22	40.36	49.26
16	xCOMET XXL	xCOMET-XXL	T1-B	33.15	59.42	44.84	51.11
17	Vertical_2	Gemma 3 1B IT	T2	30.24	46.64	70.22	56.05
18	FACET	GPT-4.1 Mini (2025-04-14)	T1	27.21	60.19	31.73	41.55
18	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	27.21	60.19	31.73	41.55
20	FACET	GPT-4.1 Mini (2025-04-14)	T2	25.70	58.05	31.73	41.03
20	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	25.70	58.05	31.73	41.03
20	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	25.70	58.05	31.73	41.03
20	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	25.70	58.05	31.73	41.03
24	gemba-poly	Llama 3.3 70B	T2	23.98	49.76	44.20	46.82
25	First try [871721]	Qwen3.6-27B	T3	23.78	55.79	31.18	40.00
26	Vertical_1	Gemma 3 1B IT	T2	20.93	43.87	55.41	48.97
27	First try [874374]	Babel-83B-Chat	T3	20.70	36.40	94.29	52.52
28	PTT-MULTITAN-v2 [875519]	—	T3	11.68	35.01	84.05	49.43
29	xCOMET XL	xCOMET-XL	T1-B	11.57	54.66	9.18	15.72
30	bytepop_pro9	Ensemble <sup>E3</sup>	T2	10.76	32.88	99.75	49.46
30	bytepop_pro10	Ensemble <sup>E4</sup>	T2	10.76	32.88	99.75	49.46
32	PTT-MULTITAN-v2 [875520]	—	T3	1.29	32.60	52.75	40.30

Table 71: WMT26 Task 3 results for eng-deu (eng\_Latn--deu\_Latn). 459/459 items (100.0%) and 6,839/13,770 item-system pairs (49.7%) are annotated, giving 7,383 human assessments (2,364 error-free, 5,019 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E2</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	47.72	53.61	67.63	59.81
2	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	47.17	56.61	61.41	58.91
3	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	46.96	50.97	71.03	59.35
4	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	45.55	52.61	64.92	58.12
5	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	45.47	51.39	67.05	58.19
6	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	44.27	51.23	64.80	57.22
7	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E1</sup>	T3	42.42	51.96	59.50	55.48
8	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T3	42.35	50.78	61.52	55.64
9	Austriadraco	GPT-5.4	T3	41.99	43.37	77.25	55.55
10	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	41.49	54.93	52.65	53.76
11	Austriadactylus	Ensemble <sup>E3</sup>	T2	40.99	41.44	80.13	54.62
11	Pterodactyl	GPT-5.4	T2	40.99	41.44	80.13	54.62
13	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	39.99	57.81	45.62	51.00
14	Cohere CAT+ ensemble	Command A+	T2	39.76	59.71	42.86	49.90
15	Wukongopterus	GPT-5.4	T3	38.55	37.13	87.33	52.11
16	cuni-v14	Gemma 4 31B	T1	38.30	43.44	68.49	53.16
17	PhraseT3-02	Gemini 3.5 Flash	T3	37.97	37.96	83.41	52.17
18	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	37.87	62.64	36.41	46.05
19	cuni-cat-v4	Gemma 4 31B	T1	37.47	43.21	67.11	52.57
20	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	37.44	42.21	69.64	52.57
21	Cohere CAT+	Command A+	T1	36.37	46.73	56.34	51.08
22	PhraseT3-01	Gemini 3.5 Flash	T3	34.91	34.03	89.57	49.33
23	AEGIS	Ensemble <sup>E4</sup>	T2	32.46	56.51	33.24	41.86
24	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T2	32.37	33.00	87.62	47.94
25	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E1</sup>	T2	31.86	33.15	85.83	47.83
26	First try [871721]	Qwen3.6-27B	T3	31.59	49.47	40.67	44.64
27	FACET	GPT-4.1 Mini (2025-04-14)	T1	27.82	47.41	35.83	40.81
28	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	27.81	47.44	35.77	40.79
29	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	27.64	46.69	36.58	41.02
29	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	27.64	46.69	36.58	41.02
31	FACET	GPT-4.1 Mini (2025-04-14)	T2	27.57	46.56	36.64	41.01
31	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	27.57	46.56	36.64	41.01
33	XCOMOffset	—	T1	26.52	37.86	53.92	44.49
34	gemba-poly	Llama 3.3 70B	T2	24.45	41.49	38.77	40.08
35	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T1	22.38	27.86	87.33	42.25
36	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E1</sup>	T1	20.93	27.87	83.35	41.77
37	xCOMET XXL	xCOMET-XXL	T1-B	20.67	38.39	35.89	37.09
38	bytepop_pro10	Ensemble <sup>E5</sup>	T2	18.96	24.30	99.88	39.09
39	bytepop_pro9	Ensemble <sup>E6</sup>	T2	17.91	23.98	99.94	38.68
40	First try [874374]	Babel-83B-Chat	T3	14.83	24.92	87.67	38.81
41	Qwen-based Omission	Qwen3-32B	T1	13.39	27.70	56.28	37.13
42	Rule-based Omission	Qwen3-32B	T1	13.28	27.62	56.39	37.08
43	xCOMET XL	xCOMET-XL	T1-B	10.56	35.94	13.77	19.91
44	RAGthoven Error-Free	Ensemble <sup>E7</sup>	T3	7.84	25.82	42.34	32.08
45	Vertical_2	Gemma 3 1B IT	T2	3.51	31.07	3.17	5.75
46	Vertical_1	Gemma 3 1B IT	T2	2.07	26.70	3.17	5.66

Table 72: WMT26 Task 3 results for eng-ekk (eng\_Latn--ekk\_Latn). 356/356 items (100.0%) and 7,235/10,680 item–system pairs (67.7%) are annotated, giving 8,094 human assessments (1,736 error-free, 6,358 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Lexicala-QE-Ensemble (STAPLE): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E2</sup> Lexicala-QE-Ensemble (Majority-Vote): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E3</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E4</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E5</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E6</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E7</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Austriadraco	GPT-5.4	T3	33.96	32.88	73.38	45.41
2	Austriadactylus	Ensemble <sup>E1</sup>	T2	33.46	31.61	76.49	44.73
2	Pterodactyl	GPT-5.4	T2	33.46	31.61	76.49	44.73
4	Wukongopterus	GPT-5.4	T3	31.90	28.54	84.17	42.62
5	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	28.69	36.83	47.48	41.49
6	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	25.79	40.56	33.21	36.52
7	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	25.35	39.93	33.21	36.26
8	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	24.26	34.95	40.16	37.38
8	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	24.26	34.95	40.16	37.38
10	AEGIS	Ensemble <sup>E2</sup>	T2	23.16	39.17	29.46	33.63
11	bytepop_pro9	Ensemble <sup>E3</sup>	T2	21.43	20.19	99.73	33.57
11	bytepop_pro10	Ensemble <sup>E4</sup>	T2	21.43	20.19	99.73	33.57
13	gemba-poly	Llama 3.3 70B	T2	20.17	27.77	50.50	35.83
14	Vertical_2	Gemma 3 1B IT	T2	19.06	28.19	44.83	34.62
15	First try [871721]	Qwen3.6-27B	T3	18.82	38.06	21.59	27.55
16	Vertical_1	Gemma 3 1B IT	T2	18.25	29.78	36.51	32.80
17	xCOMET XXL	xCOMET-XXL	T1-B	15.74	29.67	28.91	29.29
18	First try [874374]	Babel-83B-Chat	T3	13.47	19.44	85.09	31.65
19	Cohere CAT+ ensemble	Command A+	T2	13.32	39.65	10.34	16.40
20	FACET	GPT-4.1 Mini (2025-04-14)	T1	11.35	33.00	11.99	17.58
20	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	11.35	33.00	11.99	17.58
22	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	10.80	31.18	12.81	18.16
23	Cohere CAT+	Command A+	T1	10.13	26.05	20.49	22.94
24	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	9.22	19.33	65.97	29.90
25	FACET	GPT-4.1 Mini (2025-04-14)	T2	7.86	26.49	12.17	16.68
25	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	7.86	26.49	12.17	16.68
25	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	7.86	26.49	12.17	16.68
25	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	7.86	26.49	12.17	16.68
29	xCOMET XL	xCOMET-XL	T1-B	4.28	20.54	15.28	17.52
30	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	2.60	16.56	93.87	28.16

Table 73: WMT26 Task 3 results for eng-hye (eng\_Latn--hye\_Armn). 364/364 items (100.0%) and 6,123/10,192 item-system pairs (60.1%) are annotated, giving 6,711 human assessments (1,093 error-free, 5,618 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	33.39	63.24	55.05	58.86
2	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	32.96	66.57	47.05	55.14
3	PhraseT3-01	Gemini 3.5 Flash	T3	32.51	54.42	78.80	64.38
4	cuni-cat-v4	Gemma 4 31B	T1	32.39	59.08	63.77	61.33
5	Wukongopterus	GPT-5.4	T3	31.91	54.72	76.59	63.83
6	cuni-v14	Gemma 4 31B	T1	31.88	58.72	63.69	61.10
7	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E1</sup>	T1	31.83	59.96	60.04	60.00
8	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	31.29	66.68	43.47	52.63
9	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E1</sup>	T2	31.25	57.70	65.37	61.30
10	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E2</sup>	T2	31.14	59.77	59.04	59.40
11	Austriadactylus	Ensemble <sup>E3</sup>	T2	30.86	57.21	66.05	61.32
11	Pterodactyl	GPT-5.4	T2	30.86	57.21	66.05	61.32
13	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	30.51	70.81	35.29	47.10
14	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E1</sup>	T3	30.33	65.83	43.09	52.08
15	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E2</sup>	T1	30.08	64.13	45.97	53.55
16	Austriadraco	GPT-5.4	T3	29.96	57.92	61.80	59.80
17	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E2</sup>	T3	29.44	68.08	37.35	48.24
18	PhraseT3-02	Gemini 3.5 Flash	T3	29.27	54.32	72.38	62.07
19	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	27.25	68.63	32.26	43.89
20	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	26.31	65.63	35.09	45.73
21	Cohere CAT+	Command A+	T1	26.14	66.40	33.47	44.50
22	AEGIS	Ensemble <sup>E4</sup>	T2	25.76	70.06	27.62	39.61
23	gemba-poly	Llama 3.3 70B	T2	23.98	58.32	45.85	51.34
24	Cohere CAT+ ensemble	Command A+	T2	23.86	77.37	17.96	29.15
25	First try [874374]	Babel-83B-Chat	T3	23.17	48.44	84.23	61.51
26	xCOMET XXL	xCOMET-XXL	T1-B	23.07	63.23	32.36	42.82
27	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	22.97	61.02	36.57	45.73
28	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	22.61	61.36	34.97	44.55
29	First try [871721]	Qwen3.6-27B	T3	22.42	68.01	24.03	35.51
30	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	21.47	71.35	19.02	30.03
31	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	21.09	64.29	26.37	37.40
32	bytepop_pro9	Ensemble <sup>E5</sup>	T2	18.65	43.38	99.90	60.49
32	bytepop_pro10	Ensemble <sup>E6</sup>	T2	18.65	43.38	99.90	60.49
34	Vertical_2	Gemma 3 1B IT	T2	16.78	62.79	19.58	29.85
35	FACET	GPT-4.1 Mini (2025-04-14)	T1	15.82	68.83	12.34	20.93
35	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	15.82	68.83	12.34	20.93
37	Qwen-based Omission	Qwen3-32B	T1	15.76	50.49	50.88	50.68
38	Rule-based Omission	Qwen3-32B	T1	15.59	50.36	50.94	50.65
39	RAGthoven Error-Free	Ensemble <sup>E7</sup>	T3	14.71	50.26	48.04	49.12
40	FACET	GPT-4.1 Mini (2025-04-14)	T2	12.12	61.44	12.00	20.08
40	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	12.12	61.44	12.00	20.08
40	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	12.12	61.44	12.00	20.08
40	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	12.12	61.44	12.00	20.08
44	Vertical_1	Gemma 3 1B IT	T2	11.25	55.12	18.46	27.65
45	xCOMET XL	xCOMET-XL	T1-B	9.89	57.17	11.90	19.70

Table 74: WMT26 Task 3 results for eng-ind (eng\_Latn--ind\_Latn). 331/331 items (100.0%) and 10,261/10,261 item-system pairs (100.0%) are annotated, giving 12,084 human assessments (4,990 error-free, 7,094 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Lexicala-QE-Ensemble (Majority-Vote): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E2</sup> Lexicala-QE-Ensemble (STAPLE): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E3</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E4</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E5</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E6</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E7</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Austriadactylus	Ensemble <sup>E1</sup>	T2	43.46	50.08	69.31	58.15
1	Pterodactyl	GPT-5.4	T2	43.46	50.08	69.31	58.15
3	Austriadraco	GPT-5.4	T3	42.81	51.45	65.06	57.46
4	Wukongopterus	GPT-5.4	T3	42.61	44.92	79.60	57.43
5	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	39.43	56.31	49.48	52.68
6	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	39.10	57.51	47.07	51.77
7	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	38.77	57.34	46.67	51.46
8	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	37.85	54.13	49.71	51.83
9	cuni-cat-v4	Gemma 4 31B	T1	37.75	48.22	60.75	53.76
10	cuni-v14	Gemma 4 31B	T1	37.64	47.93	61.15	53.74
11	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	37.57	58.67	42.59	49.35
12	PhraseT3-01	Gemini 3.5 Flash	T3	36.32	39.93	79.66	53.20
13	PhraseT3-02	Gemini 3.5 Flash	T3	36.11	41.58	74.08	53.26
14	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	34.62	58.92	36.84	45.33
15	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	34.57	48.65	52.64	50.57
16	AEGIS	Ensemble <sup>E2</sup>	T2	34.46	67.07	28.91	40.40
17	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	33.47	58.65	35.06	43.88
18	First try [871721]	Qwen3.6-27B	T3	29.29	55.36	31.15	39.87
19	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	27.99	55.68	28.45	37.66
20	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	26.63	64.04	19.66	30.08
21	Vertical_2	Gemma 3 1B IT	T2	26.42	61.23	21.15	31.44
22	Cohere CAT+	Command A+	T1	25.79	48.01	34.02	39.83
23	Cohere CAT+ ensemble	Command A+	T2	24.36	62.10	17.70	27.55
24	bytepop_pro10	Ensemble <sup>E3</sup>	T2	24.10	28.01	100.00	43.76
25	gemba-poly	Llama 3.3 70B	T2	23.85	48.65	28.91	36.27
26	bytepop_pro9	Ensemble <sup>E4</sup>	T2	23.43	27.76	100.00	43.46
27	RAGthoven Error-Free	Ensemble <sup>E5</sup>	T3	23.12	37.11	54.20	44.06
28	xCOMET XXL	xCOMET-XXL	T1-B	21.84	44.89	30.29	36.17
29	First try [874374]	Babel-83B-Chat	T3	17.49	27.74	89.43	42.35
30	FACET	GPT-4.1 Mini (2025-04-14)	T1	15.35	47.79	13.68	21.27
30	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	15.35	47.79	13.68	21.27
32	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	12.19	40.86	14.25	21.13
32	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	12.19	40.86	14.25	21.13
34	FACET	GPT-4.1 Mini (2025-04-14)	T2	12.12	40.72	14.25	21.12
34	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	12.12	40.72	14.25	21.12
36	Qwen-based Omission	Qwen3-32B	T1	10.22	28.60	51.72	36.83
37	Rule-based Omission	Qwen3-32B	T1	10.19	28.58	51.78	36.83
38	xCOMET XL	xCOMET-XL	T1-B	8.69	35.84	12.59	18.63
39	Vertical_1	Gemma 3 1B IT	T2	8.60	48.99	4.20	7.73

Table 75: WMT26 Task 3 results for eng-isl (eng\_Latn--isl\_Latn). 360/360 items (100.0%) and 6,823/11,160 item-system pairs (61.1%) are annotated, giving 7,383 human assessments (1,740 error-free, 5,643 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E5</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Austriadraco	GPT-5.4	T3	39.57	56.08	70.62	62.51
2	Austriadactylus	Ensemble <sup>E1</sup>	T2	39.28	54.48	74.07	62.79
2	Pterodactyl	GPT-5.4	T2	39.28	54.48	74.07	62.79
4	Wukongopterus	GPT-5.4	T3	38.27	50.34	83.50	62.81
5	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	38.19	61.31	56.22	58.66
6	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	37.69	56.68	65.30	60.69
7	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	36.04	62.84	49.06	55.10
7	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	36.04	62.84	49.06	55.10
9	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	35.93	65.70	44.19	52.84
10	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	34.72	64.45	43.81	52.16
11	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	31.57	63.43	39.25	48.50
12	First try [871721]	Qwen3.6-27B	T3	26.89	66.93	26.27	37.73
13	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	24.75	71.12	19.42	30.51
14	Cohere CAT+ ensemble	Command A+	T2	24.57	66.82	22.53	33.70
15	AEGIS	Ensemble <sup>E2</sup>	T2	24.39	61.34	28.22	38.66
16	bytepop_pro9	Ensemble <sup>E3</sup>	T2	22.15	38.00	99.72	55.03
16	bytepop_pro10	Ensemble <sup>E4</sup>	T2	22.15	38.00	99.72	55.03
18	Vertical_2	Gemma 3 1B IT	T2	21.31	53.37	35.04	42.31
19	Cohere CAT+	Command A+	T1	20.14	51.24	37.15	43.07
20	gemba-poly	Llama 3.3 70B	T2	19.44	54.04	29.01	37.75
21	First try [874374]	Babel-83B-Chat	T3	17.97	40.67	78.79	53.65
22	FACET	GPT-4.1 Mini (2025-04-14)	T1	17.47	59.85	17.00	26.48
22	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	17.47	59.85	17.00	26.48
24	FACET	GPT-4.1 Mini (2025-04-14)	T2	16.00	57.36	16.78	25.97
24	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	16.00	57.36	16.78	25.97
24	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	16.00	57.36	16.78	25.97
24	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	16.00	57.36	16.78	25.97
28	Vertical_1	Gemma 3 1B IT	T2	13.12	51.92	17.38	26.04
29	xCOMET XXL	xCOMET-XXL	T1-B	12.37	46.19	27.66	34.60
30	xCOMET XL	xCOMET-XL	T1-B	5.62	41.58	15.74	22.84

Table 76: WMT26 Task 3 results for eng-kaz (eng\_Latn--kaz\_Cyrl). 438/438 items (100.0%) and 7,981/12,702 item-system pairs (62.8%) are annotated, giving 9,171 human assessments (3,182 error-free, 5,989 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	AEGIS	Ensemble <sup>E1</sup>	T2	32.50	68.66	17.92	28.42
2	Cohere CAT+ ensemble	Command A+	T2	26.41	36.24	28.05	31.63
3	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	23.13	24.78	39.48	30.45
4	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	22.96	22.93	44.29	30.22
5	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	22.25	25.44	35.32	29.58
6	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	19.46	32.50	18.57	23.64
7	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	16.98	18.56	39.48	25.25
8	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	16.78	18.54	38.83	25.09
9	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	15.84	21.14	27.01	23.72
10	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	15.58	15.88	47.92	23.85
11	FACET	GPT-4.1 Mini (2025-04-14)	T1	15.51	23.28	21.56	22.39
12	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	15.48	23.31	21.43	22.33
13	bytepop_pro9	Ensemble <sup>E2</sup>	T2	14.50	11.19	90.91	19.93
13	bytepop_pro10	Ensemble <sup>E3</sup>	T2	14.50	11.19	90.91	19.93
15	Vertical_2	Gemma 3 1B IT	T2	14.47	41.67	7.14	12.20
16	FACET	GPT-4.1 Mini (2025-04-14)	T2	13.19	20.15	21.43	20.77
16	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	13.19	20.15	21.43	20.77
18	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	13.17	20.17	21.30	20.72
18	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	13.17	20.17	21.30	20.72
20	Austriadraco	GPT-5.4	T3	12.60	14.29	45.06	21.69
21	Pterodactyl	GPT-5.4	T2	12.53	13.67	50.13	21.49
22	Austriadactylus	Ensemble <sup>E4</sup>	T2	12.20	11.96	68.83	20.37
23	First try [874374]	Babel-83B-Chat	T3	12.17	16.74	28.44	21.08
24	Wukongopterus	GPT-5.4	T3	10.83	11.81	62.73	19.88
25	First try [871721]	Qwen3.6-27B	T3	9.30	16.57	18.18	17.34
26	Cohere CAT+	Command A+	T1	8.74	13.12	33.38	18.83
27	Vertical_1	Gemma 3 1B IT	T2	8.15	30.00	3.90	6.90
28	gemba-poly	Llama 3.3 70B	T2	1.35	9.33	17.92	12.27
29	xCOMET XL	xCOMET-XL	T1-B	-0.97	7.92	18.05	11.01
30	xCOMET XXL	xCOMET-XXL	T1-B	-1.92	7.72	30.13	12.28

Table 77: WMT26 Task 3 results for eng-lij (eng\_Latn--lij\_Latn). 411/411 items (100.0%) and 8,601/11,919 item-system pairs (72.2%) are annotated, giving 9,084 human assessments (770 error-free, 8,314 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E2</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E3</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	AEGIS	Ensemble <sup>E1</sup>	T2	20.33	67.44	7.07	12.80
2	bytepop_pro9	Ensemble <sup>E2</sup>	T2	15.39	9.45	99.27	17.26
2	bytepop_pro10	Ensemble <sup>E3</sup>	T2	15.39	9.45	99.27	17.26
4	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	15.26	18.78	27.07	22.18
5	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	15.14	20.73	22.20	21.44
6	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	13.69	15.97	30.98	21.08
7	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	13.61	15.84	31.22	21.02
8	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	12.29	22.71	12.68	16.28
9	Austriadactylus	Ensemble <sup>E4</sup>	T2	11.47	10.52	64.63	18.09
10	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	11.04	19.78	13.41	15.99
11	Wukongopterus	GPT-5.4	T3	10.95	10.67	59.02	18.07
12	Pterodactyl	GPT-5.4	T2	9.58	11.63	38.29	17.84
13	Austriadraco	GPT-5.4	T3	8.69	11.76	31.71	17.16
14	Cohere CAT+	Command A+	T1	8.31	12.19	26.34	16.67
15	Vertical_1	Gemma 3 1B IT	T2	6.98	33.33	2.20	4.12
16	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	6.37	16.38	7.07	9.88
17	FACET	GPT-4.1 Mini (2025-04-14)	T1	6.33	16.29	7.07	9.86
18	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	6.07	11.40	19.27	14.32
19	Vertical_2	Gemma 3 1B IT	T2	5.73	44.44	0.98	1.91
20	Cohere CAT+ ensemble	Command A+	T2	3.68	15.71	2.68	4.58
21	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	3.22	10.87	7.32	8.75
21	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	3.22	10.87	7.32	8.75
23	FACET	GPT-4.1 Mini (2025-04-14)	T2	3.19	10.83	7.32	8.73
23	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	3.19	10.83	7.32	8.73
25	xCOMET XXL	xCOMET-XXL	T1-B	3.15	7.85	65.61	14.02
26	First try [874374]	Babel-83B-Chat	T3	1.12	8.58	4.88	6.22
27	gemba-poly	Llama 3.3 70B	T2	0.90	7.99	8.54	8.25
28	First try [871721]	Qwen3.6-27B	T3	0.42	7.84	2.93	4.26
29	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	-0.11	6.94	1.22	2.07
30	xCOMET XL	xCOMET-XL	T1-B	-1.27	6.68	27.32	10.73

Table 78: WMT26 Task 3 results for eng-lld (eng\_Latn--lld\_Latn). 376/376 items (100.0%) and 5,201/10,152 item-system pairs (51.2%) are annotated, giving 5,706 human assessments (410 error-free, 5,296 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E2</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E3</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	38.12	59.62	58.62	59.12
2	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	37.45	53.11	72.87	61.44
3	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	36.45	59.01	56.45	57.70
4	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	34.66	55.51	60.64	57.96
5	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	34.55	56.23	58.65	57.41
6	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	34.33	59.08	51.85	55.23
7	Austriadraco	GPT-5.4	T3	34.15	49.01	78.13	60.24
8	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	33.88	67.22	37.67	48.28
9	Cohere CAT+ ensemble	Command A+	T2	33.87	69.65	34.74	46.36
10	Austriadactylus	Ensemble <sup>E1</sup>	T2	33.75	48.02	80.56	60.17
10	Pterodactyl	GPT-5.4	T2	33.75	48.02	80.56	60.17
12	Wukongopterus	GPT-5.4	T3	32.49	45.51	86.57	59.66
13	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	30.98	54.91	53.84	54.37
14	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	29.39	64.15	33.44	43.96
15	FACET	GPT-4.1 Mini (2025-04-14)	T1	29.37	64.12	33.44	43.96
16	Cohere CAT+	Command A+	T1	29.11	62.92	34.58	44.63
17	First try [871721]	Qwen3.6-27B	T3	28.92	58.86	40.74	48.15
18	FACET	GPT-4.1 Mini (2025-04-14)	T2	26.96	60.63	33.67	43.29
18	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	26.96	60.63	33.67	43.29
18	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	26.96	60.63	33.67	43.29
18	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	26.96	60.63	33.67	43.29
22	XCOMOffset	—	T1	26.80	55.68	42.29	48.07
23	AEGIS	Ensemble <sup>E2</sup>	T2	26.37	59.09	34.88	43.87
24	Vertical_2	Gemma 3 1B IT	T2	23.78	59.87	28.36	38.49
25	First try [874374]	Babel-83B-Chat	T3	22.39	39.19	93.96	55.32
26	gemba-poly	Llama 3.3 70B	T2	22.01	50.58	42.80	46.36
27	xCOMET XXL	xCOMET-XXL	T1-B	16.69	45.95	42.06	43.92
28	PTT-MULTITAN-v2 [875519]	—	T3	14.52	38.28	83.05	52.41
29	bytepop_pro9	Ensemble <sup>E3</sup>	T2	13.81	35.53	99.84	52.41
29	bytepop_pro10	Ensemble <sup>E4</sup>	T2	13.81	35.53	99.84	52.41
31	xCOMET XL	xCOMET-XL	T1-B	13.59	50.43	20.15	28.80
32	PTT-MULTITAN-v2 [875520]	—	T3	10.76	39.23	58.44	46.95
33	Vertical_1	Gemma 3 1B IT	T2	8.91	48.97	10.86	17.78

Table 79: WMT26 Task 3 results for eng-rus (eng\_Latn--rus\_Cyrl). 426/426 items (100.0%) and 11,131/13,206 item-system pairs (84.3%) are annotated, giving 12,774 human assessments (4,372 error-free, 8,402 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	35.11	34.50	43.07	38.31
2	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	34.62	27.57	54.74	36.67
3	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	34.43	27.97	53.28	36.68
4	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E1</sup>	T3	33.30	39.82	32.85	36.00
5	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	31.49	21.39	62.77	31.91
6	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T3	31.44	35.66	33.58	34.59
7	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T1	30.08	23.96	50.36	32.47
8	cuni-v14	Gemma 4 31B	T1	29.68	24.35	48.18	32.35
9	Austriadraco	GPT-5.4	T3	28.96	21.25	54.74	30.61
10	Austriadactylus	Ensemble <sup>E3</sup>	T2	28.47	14.99	81.75	25.34
11	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	28.44	27.66	37.96	32.00
12	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	28.31	21.74	51.09	30.50
12	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	28.31	21.74	51.09	30.50
14	cuni-cat-v4	Gemma 4 31B	T1	28.27	24.49	43.80	31.41
15	Wukongopterus	GPT-5.4	T3	27.94	15.22	78.10	25.48
16	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E1</sup>	T1	27.91	21.79	49.64	30.29
17	Pterodactyl	GPT-5.4	T2	27.53	18.66	59.12	28.37
18	Cohere CAT+ ensemble	Command A+	T2	26.62	52.50	15.33	23.73
19	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T2	26.53	19.67	51.82	28.51
20	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E1</sup>	T2	23.10	16.15	53.28	24.79
21	bytepop_pro9	Ensemble <sup>E4</sup>	T2	22.33	9.50	100.00	17.35
21	bytepop_pro10	Ensemble <sup>E5</sup>	T2	22.33	9.50	100.00	17.35
23	AEGIS	Ensemble <sup>E6</sup>	T2	22.08	45.95	12.41	19.54
24	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	21.92	30.43	20.44	24.45
25	PhraseT3-01	Gemini 3.5 Flash	T3	21.05	11.21	76.64	19.55
26	First try [871721]	Qwen3.6-27B	T3	19.55	26.47	19.71	22.59
27	PhraseT3-02	Gemini 3.5 Flash	T3	19.31	10.90	70.80	18.89
28	Cohere CAT+	Command A+	T1	16.23	13.42	38.69	19.92
29	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	13.76	13.54	28.47	18.35
30	RAGthoven Error-Free	Ensemble <sup>E7</sup>	T3	12.81	11.98	31.39	17.34
31	FACET	GPT-4.1 Mini (2025-04-14)	T2	11.41	14.63	17.52	15.95
31	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	11.41	14.63	17.52	15.95
31	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	11.41	14.63	17.52	15.95
31	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	11.41	14.63	17.52	15.95
35	FACET	GPT-4.1 Mini (2025-04-14)	T1	10.85	14.57	16.06	15.28
35	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	10.85	14.57	16.06	15.28
37	First try [874374]	Babel-83B-Chat	T3	10.74	7.80	59.12	13.78
38	Vertical_1	Gemma 3 1B IT	T2	9.20	36.36	2.92	5.41
39	gemba-poly	Llama 3.3 70B	T2	7.07	14.29	7.30	9.66
40	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	3.81	9.86	5.11	6.73
41	Vertical_2	Gemma 3 1B IT	T2	3.66	13.64	2.19	3.77
42	Qwen-based Omission	Qwen3-32B	T1	-0.19	4.71	51.09	8.63
43	Rule-based Omission	Qwen3-32B	T1	-0.20	4.71	51.09	8.63
44	xCOMET XXL	xCOMET-XXL	T1-B	-5.23	4.14	67.15	7.81
45	xCOMET XL	xCOMET-XL	T1-B	-7.07	2.73	20.44	4.82

Table 80: WMT26 Task 3 results for eng-sme (eng\_Latn--sme\_Latn). 165/493 items (33.5%) and 2,532/13,311 item-system pairs (19.0%) are annotated, giving 2,883 human assessments (137 error-free, 2,746 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Lexicala-QE-Ensemble (STAPLE): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E2</sup> Lexicala-QE-Ensemble (Majority-Vote): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E3</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E4</sup> byt Pop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E5</sup> byt Pop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E6</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E7</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	49.37	60.88	70.81	65.47
2	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	48.45	55.68	79.34	65.44
3	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	47.82	63.92	62.69	63.30
4	Austriadraco	GPT-5.4	T3	44.86	51.53	81.83	63.24
5	Austriadactylus	Ensemble <sup>E1</sup>	T2	44.31	49.86	84.85	62.81
5	Pterodactyl	GPT-5.4	T2	44.31	49.86	84.85	62.81
7	Wukongopterus	GPT-5.4	T3	42.06	45.81	91.19	60.99
8	Cohere CAT+ ensemble	Command A+	T2	41.92	74.11	39.36	51.41
9	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	41.54	56.74	63.56	59.96
9	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	41.54	56.74	63.56	59.96
11	AEGIS	Ensemble <sup>E2</sup>	T2	39.25	62.64	48.59	54.73
12	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	38.70	67.96	40.56	50.80
13	Cohere CAT+	Command A+	T1	37.48	55.41	57.85	56.60
14	FACET	GPT-4.1 Mini (2025-04-14)	T1	34.81	67.56	34.38	45.57
14	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	34.81	67.56	34.38	45.57
16	FACET	GPT-4.1 Mini (2025-04-14)	T2	27.89	56.92	34.84	43.23
16	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	27.89	56.92	34.84	43.23
16	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	27.89	56.92	34.84	43.23
16	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	27.89	56.92	34.84	43.23
20	gamba-poly	Llama 3.3 70B	T2	25.16	50.38	40.64	44.99
21	xCOMET XXL	xCOMET-XXL	T1-B	23.98	49.30	40.14	44.25
22	bytepop_pro9	Ensemble <sup>E3</sup>	T2	23.71	33.90	100.00	50.63
22	bytepop_pro10	Ensemble <sup>E4</sup>	T2	23.71	33.90	100.00	50.63
24	First try [874374]	Babel-83B-Chat	T3	22.35	34.89	94.56	50.97
25	Vertical_2	Gemma 3 1B IT	T2	19.54	55.81	19.95	29.39
26	Vertical_1	Gemma 3 1B IT	T2	13.78	43.55	25.80	32.40
27	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	12.21	44.12	19.89	27.42
28	First try [871721]	Qwen3.6-27B	T3	11.75	90.97	2.35	4.57
29	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	11.68	49.73	11.33	18.46
30	xCOMET XL	xCOMET-XL	T1-B	9.61	43.12	14.49	21.69

Table 81: WMT26 Task 3 results for eng-tha (eng\_Latn--tha\_Thai). 451/451 items (100.0%) and 13,981/13,981 item-system pairs (100.0%) are annotated, giving 18,639 human assessments (5,585 error-free, 13,054 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	44.59	65.19	69.03	67.06
2	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	42.82	71.60	54.04	61.59
3	Austriadraco	GPT-5.4	T3	41.83	60.91	73.49	66.61
4	Austriadactylus	Ensemble <sup>E1</sup>	T2	41.73	59.68	76.30	66.97
4	Pterodactyl	GPT-5.4	T2	41.73	59.68	76.30	66.97
6	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	40.52	71.11	50.71	59.20
7	Wukongopterus	GPT-5.4	T3	40.48	56.20	83.02	67.03
8	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	39.19	72.08	46.85	56.79
9	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	38.09	66.81	53.63	59.50
10	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	37.99	65.77	55.43	60.16
11	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	37.69	66.36	53.71	59.37
12	Cohere CAT+ ensemble	Command A+	T2	37.56	76.30	38.47	51.15
13	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	36.08	76.48	35.86	48.83
14	Cohere CAT+	Command A+	T1	32.82	67.53	42.16	51.91
15	First try [871721]	Qwen3.6-27B	T3	32.40	71.27	35.88	47.73
16	FACET	GPT-4.1 Mini (2025-04-14)	T1	27.56	71.16	27.60	39.77
16	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	27.56	71.16	27.60	39.77
18	First try [874374]	Babel-83B-Chat	T3	26.18	46.09	92.13	61.44
19	AEGIS	Ensemble <sup>E2</sup>	T2	25.33	70.73	24.30	36.18
20	xCOMET XXL	xCOMET-XXL	T1-B	25.25	59.98	40.40	48.28
21	FACET	GPT-4.1 Mini (2025-04-14)	T2	24.18	66.03	27.66	38.98
21	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	24.18	66.03	27.66	38.98
23	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	24.16	66.01	27.64	38.96
23	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	24.16	66.01	27.64	38.96
25	gemma-poly	Llama 3.3 70B	T2	22.60	58.37	37.65	45.78
26	PTT-MULTITAN-v2 [875519]	—	T3	20.36	45.62	83.65	59.04
27	bytepop_pro9	Ensemble <sup>E3</sup>	T2	19.55	41.95	99.59	59.03
27	bytepop_pro10	Ensemble <sup>E4</sup>	T2	19.55	41.95	99.59	59.03
29	XCOMOffset	—	T1	18.57	54.74	36.21	43.59
30	PTT-MULTITAN-v2 [875520]	—	T3	14.84	47.70	52.84	50.14
31	Vertical_1	Gemma 3 1B IT	T2	13.35	47.90	45.51	46.67
32	Vertical_2	Gemma 3 1B IT	T2	11.74	48.97	33.21	39.58
33	xCOMET XL	xCOMET-XL	T1-B	10.20	54.67	13.47	21.61

Table 82: WMT26 Task 3 results for eng-ukr (eng\_Latn--ukr\_Cyrl). 406/406 items (100.0%) and 10,860/12,180 item-system pairs (89.2%) are annotated, giving 12,999 human assessments (5,131 error-free, 7,868 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	37.93	40.40	72.21	51.81
2	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	36.30	41.07	66.37	50.74
3	Cohere CAT+ ensemble	Command A+	T2	34.01	47.18	47.43	47.31
4	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	33.00	33.67	81.59	47.67
5	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	32.84	38.10	66.02	48.32
6	Cohere CAT+	Command A+	T1	31.51	38.71	60.71	47.28
7	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	31.34	39.11	59.12	47.08
8	cuni-cat-v4	Gemma 4 31B	T1	30.21	31.18	84.42	45.54
9	Austriadracο	GPT-5.4	T3	30.13	29.30	92.21	44.47
10	Austriadactylus	Ensemble <sup>E1</sup>	T2	29.88	28.75	93.98	44.03
10	Pterodactyl	GPT-5.4	T2	29.88	28.75	93.98	44.03
12	cuni-v14	Gemma 4 31B	T1	29.66	30.89	84.25	45.20
13	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	27.42	43.30	39.47	41.30
13	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	27.42	43.30	39.47	41.30
15	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	27.21	35.77	58.05	44.26
16	Wukongopterus	GPT-5.4	T3	25.46	26.37	95.22	41.31
17	FACET	GPT-4.1 Mini (2025-04-14)	T2	25.17	35.31	53.63	42.59
17	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	25.17	35.31	53.63	42.59
17	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	25.17	35.31	53.63	42.59
17	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	25.17	35.31	53.63	42.59
21	FACET	GPT-4.1 Mini (2025-04-14)	T1	25.01	35.38	52.92	42.41
21	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	25.01	35.38	52.92	42.41
23	xCOMET XXL	xCOMET-XXL	T1-B	23.94	39.12	39.47	39.30
24	AEGIS	Ensemble <sup>E2</sup>	T2	22.21	38.21	36.99	37.59
25	First try [871721]	Qwen3.6-27B	T3	21.28	35.57	41.24	38.20
26	xCOMET XL	xCOMET-XL	T1-B	16.28	49.25	11.68	18.88
27	gemba-poly	Llama 3.3 70B	T2	12.78	26.38	52.39	35.09
28	PTT-MULTITAN-v2 [875520]	—	T3	11.43	24.24	66.37	35.51
29	bytepop_pro10	Ensemble <sup>E3</sup>	T2	6.54	20.44	100.00	33.94
30	bytepop_pro9	Ensemble <sup>E4</sup>	T2	6.03	20.39	100.00	33.87
31	Vertical_2	Gemma 3 1B IT	T2	4.10	23.44	22.65	23.04
32	Vertical_1	Gemma 3 1B IT	T2	2.45	21.82	26.73	24.03
33	PTT-MULTITAN-v2 [875519]	—	T3	0.15	20.14	64.07	30.65
34	First try [874374]	Babel-83B-Chat	T3	−0.49	19.99	75.75	31.63

Table 83: WMT26 Task 3 results for eng-zho\_CN (eng\_Latn--zho\_Hans). 165/481 items (34.3%) and 2,388/13,949 item–system pairs (17.1%) are annotated, giving 2,811 human assessments (565 error-free, 2,246 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC <sub>×100</sub>	P(%)	R(%)	F1(%)
1	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	32.91	48.51	47.24	47.87
2	Cohere CAT+ ensemble	Command A+	T2	31.97	50.08	42.40	45.92
3	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	31.61	35.99	78.59	49.37
4	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	31.21	47.29	45.72	46.49
4	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	31.21	47.29	45.72	46.49
6	FACET	GPT-4.1 Mini (2025-04-14)	T1	30.51	43.46	52.35	47.49
6	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	30.51	43.46	52.35	47.49
8	Cohere CAT+	Command A+	T1	30.50	41.45	57.60	48.21
9	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	30.45	39.21	64.23	48.69
10	FACET	GPT-4.1 Mini (2025-04-14)	T2	30.07	42.91	52.62	47.27
10	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	30.07	42.91	52.62	47.27
10	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	30.07	42.91	52.62	47.27
10	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	30.07	42.91	52.62	47.27
14	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	29.06	39.38	59.94	47.54
15	Austriadactylus	Ensemble <sup>E1</sup>	T2	27.86	31.32	88.81	46.31
15	Pterodactyl	GPT-5.4	T2	27.86	31.32	88.81	46.31
17	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	27.40	38.73	57.46	46.27
18	cuni-cat-v4	Gemma 4 31B	T1	26.70	32.38	81.08	46.28
19	Austriadraco	GPT-5.4	T3	26.66	31.60	84.67	46.02
20	cuni-v14	Gemma 4 31B	T1	26.45	32.27	80.94	46.14
21	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	26.26	39.58	51.66	44.82
22	Wukongopterus	GPT-5.4	T3	26.21	29.55	93.23	44.88
23	AEGIS	Ensemble <sup>E2</sup>	T2	25.27	44.96	36.33	40.18
24	xCOMET XXL	xCOMET-XXL	T1-B	19.77	38.85	35.36	37.02
25	gemba-poly	Llama 3.3 70B	T2	19.13	32.97	54.42	41.06
26	First try [871721]	Qwen3.6-27B	T3	16.32	40.98	21.96	28.60
27	PTT-MULTITAN-v2 [875519]	—	T3	10.72	25.72	76.66	38.51
28	First try [874374]	Babel-83B-Chat	T3	8.34	24.53	83.01	37.87
29	bytepop_pro10	Ensemble <sup>E3</sup>	T2	8.31	23.13	100.00	37.57
30	bytepop_pro9	Ensemble <sup>E4</sup>	T2	8.02	23.09	100.00	37.52
31	PTT-MULTITAN-v2 [875520]	—	T3	6.31	25.20	56.49	34.85
32	Vertical_2	Gemma 3 1B IT	T2	6.07	28.28	20.86	24.01
33	xCOMET XL	xCOMET-XL	T1-B	4.84	32.33	5.94	10.04
34	Vertical_1	Gemma 3 1B IT	T2	−0.30	22.37	23.20	22.78

Table 84: WMT26 Task 3 results for eng-zho\_TW (eng\_Latn--zho\_Hant\_TW). 192/325 items (59.1%) and 2,880/9,100 item–system pairs (31.6%) are annotated, giving 3,204 human assessments (724 error-free, 2,480 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> AEGIS: XLM-RoBERTa-large; mT5-large; <sup>E3</sup> bytepop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E4</sup> bytepop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna.

Rank	Submission	Underlying model	Track	MCC $\times 100$	P(%)	R(%)	F1(%)
1	Wukongopterus	GPT-5.4	T3	35.59	48.89	73.55	58.74
2	Austriadactylus	Ensemble <sup>E1</sup>	T2	34.53	52.22	61.98	56.69
2	Pterodactyl	GPT-5.4	T2	34.53	52.22	61.98	56.69
4	Austriadraco	GPT-5.4	T3	34.12	53.51	57.89	55.61
5	cuni-v14	Gemma 4 31B	T1	34.06	50.39	65.81	57.08
6	cuni-cat-v4	Gemma 4 31B	T1	33.79	50.30	65.46	56.89
7	PhraseT3-01	Gemini 3.5 Flash	T3	30.08	45.36	72.41	55.78
8	Gemma 4 Reasoning	Gemma 4 31B IT	T1-B	28.89	54.36	44.11	48.70
9	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T2	28.77	52.99	46.80	49.70
10	Gemma 4 No Reasoning	Gemma 4 31B IT	T1-B	27.25	57.06	35.61	43.85
11	RAGthoven Error-Free	Ensemble <sup>E3</sup>	T3	26.88	41.32	80.81	54.68
12	PhraseT3-02	Gemini 3.5 Flash	T3	26.68	43.73	70.15	53.87
13	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E4</sup>	T2	24.93	52.31	39.16	44.79
14	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T1	24.48	52.56	37.56	43.81
15	Gemini 3.6 Flash	Gemini 3.6 Flash	T1-B	23.67	54.47	32.16	40.44
16	gemba-poly	Llama 3.3 70B	T2	23.64	48.51	45.15	46.77
17	GPT-5.6 Terra	GPT-5.6 Terra	T1-B	22.76	59.35	23.68	33.85
18	Lexicala-QE-Ensemble (Majority-Vote)	Ensemble <sup>E2</sup>	T3	22.64	57.55	25.53	35.37
19	fluencyscore-multispan-v2	Gemini 3.5 Flash	T1	22.38	51.28	35.13	41.69
20	fluencyscore-multispan-v2.1	Gemini 3.5 Flash	T1	21.46	51.35	32.72	39.97
21	Vertical_2	Gemma 3 1B IT	T2	20.19	54.17	25.05	34.26
22	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E4</sup>	T3	19.43	58.39	18.73	28.36
23	First try [871721]	Qwen3.6-27B	T3	19.40	60.80	16.65	26.14
24	fluency2-gemini35-esa	Gemini 3.5 Flash	T2	19.36	52.71	25.41	34.29
25	Vertical_1	Gemma 3 1B IT	T2	19.13	48.20	33.63	39.62
26	bytepop_pro9	Ensemble <sup>E5</sup>	T2	19.08	34.55	97.87	51.07
26	bytepop_pro10	Ensemble <sup>E6</sup>	T2	19.08	34.55	97.87	51.07
28	AEGIS	Ensemble <sup>E7</sup>	T2	18.95	55.41	20.94	30.39
29	Lexicala-QE-Ensemble (STAPLE)	Ensemble <sup>E4</sup>	T1	18.75	51.31	26.29	34.77
30	GPT-5.6 Sol Medium Reasoning	GPT-5.6 Sol	T1-B	18.72	61.58	15.05	24.19
31	First try [874374]	Babel-83B-Chat	T3	18.41	36.03	89.01	51.29
32	Qwen 3.6 Reasoning	Qwen3.6-35B-A3B	T1-B	17.64	54.01	20.00	29.19
33	Qwen-based Omission	Qwen3-32B	T1	13.51	39.09	50.05	43.90
34	Rule-based Omission	Qwen3-32B	T1	13.45	39.04	50.10	43.89
35	doc-FACET	GPT-4.1 Mini (2025-04-14)	T1	13.09	51.15	14.14	22.15
36	FACET	GPT-4.1 Mini (2025-04-14)	T1	13.08	51.10	14.16	22.18
37	FACET [869556]	GPT-4.1 Mini (2025-04-14)	T3	10.67	46.52	14.42	22.01
37	doc-FACET	GPT-4.1 Mini (2025-04-14)	T2	10.67	46.52	14.42	22.01
39	FACET	GPT-4.1 Mini (2025-04-14)	T2	10.63	46.45	14.44	22.03
39	FACET [869557]	GPT-4.1 Mini (2025-04-14)	T3	10.63	46.45	14.44	22.03
41	Cohere CAT+ ensemble	Command A+	T2	9.97	57.63	5.56	10.14
42	Qwen 3.6 No Reasoning	Qwen3.6-35B-A3B	T1-B	8.74	52.03	6.19	11.07
43	Cohere CAT+	Command A+	T1	6.46	39.53	15.23	21.99
44	XCOMOffset	—	T1	-4.21	28.17	23.50	25.63
45	xCOMET XXL	xCOMET-XXL	T1-B	-4.24	27.63	18.43	22.11
46	xCOMET XL	xCOMET-XL	T1-B	-8.13	16.71	3.27	5.48

Table 85: WMT26 Task 3 results for zho-jpn (zho\_Hans--jpn\_Jpan). 335/335 items (100.0%) and 10,385/10,385 item-system pairs (100.0%) are annotated, giving 12,528 human assessments (3,940 error-free, 8,588 error). Track: T3 is a direct Task 3 submission, T1 and T2 are conversions of a Task 1 or Task 2 submission, and T1-B is a converted organizer Task 1 baseline. Light-grey rows are organizer baselines. Ensemble components named in the model cards are <sup>E1</sup> Austriadactylus: GPT-5.4; GPT-5.6 Terra; <sup>E2</sup> Lexicala-QE-Ensemble (Majority-Vote): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E3</sup> RAGthoven Error-Free: Gemma 4 31B IT; Mistral Large 3; Gemini 3.1 Pro; <sup>E4</sup> Lexicala-QE-Ensemble (STAPLE): Gemini 2.5 Flash; DeepSeek V4 Flash; Qwen Plus; <sup>E5</sup> bytetop\_pro9: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E6</sup> bytetop\_pro10: GPT-5.4; Gemini 3.5 Flash; GPT-5.6 Luna; <sup>E7</sup> AEGIS: XLM-RoBERTa-large; mT5-large.