Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help

Alon Lavie[1], **Greg Hanneman**[2], **Sweta Agrawal**[3], **Diptesh Kanojia**[4]
Chi-kiu Lo 羅致翹[5], **Vilém Zouhar**[6], **Frédéric Blain**[7], **Chrysoula Zerva**[8,9]
Eleftherios Avramidis[10], **Sourabh Deoghare**[11], **Archchana Sindhujan**[4]
Jiayi Wang[12], **David I. Adelani**[13,14], **Brian Thompson**[2], **Tom Kocmi**[15]
Markus Freitag[3], **Daniel Deutsch**[3]
[1]Carnegie Mellon University [2]Unaffiliated [3]Google [4]University of Surrey
[5]National Research Council Canada [6]ETH Zurich [7]Tilburg University
[8]Instituto Superior Técnico, Universidade de Lisboa [9]Instituto de Telecomunicações
[10]German Research Center for Artificial Intelligence (DFKI)
[11]Indian Institute of Technology, Bombay [12]University College London
[13]McGill University [14]Mila - Quebec AI Institute [15]Cohere
wmt-qe-metrics-organizers@googlegroups.com

Abstract

The WMT25 Shared Task on Automated Translation Evaluation Systems evaluates metrics and quality estimation systems that assess the quality of language translation systems. This task unifies and consolidates the separate WMT shared tasks on Machine Translation Evaluation Metrics and Quality Estimation from previous years. Our primary goal is to encourage the development and assessment of new state-of-the-art translation quality evaluation systems. The shared task this year consisted of three subtasks: (1) segment-level quality score prediction, (2) span-level translation error annotation, and (3) quality-informed segment-level error correction. The evaluation data for the shared task were provided by the General MT shared task and were complemented by "challenge sets" from both the organizers and participants. Task 1 results indicate the strong performance of large LLMs at the system level, while reference-based baseline metrics outperform LLMs at the segment level. Task 2 results indicate that accurate error detection and balancing precision and recall are persistent challenges. Task 3 results show that minimal editing is challenging even when informed by quality indicators. Robustness across the broad diversity of languages remains a major challenge across all three subtasks.

1 Introduction

The WMT25 Shared Task on Automated Translation Quality Evaluation Systems[1] evaluates automated systems for assessing and improving translation quality, including automated quality metrics, quality estimation systems, and quality-informed translation error correction. This task builds on previous years' shared tasks (Freitag et al., 2024; Zerva et al., 2024) and unifies previously separate WMT shared tasks on Machine Translation Evaluation Metrics and Quality Estimation. Automated translation quality evaluation systems play a critical role in the research, development and deployment of machine translation systems, and more recently, of multilingual LLMs. They are also critical components in automated translation workflows for large-scale commercial translation use-cases.

Source	Non parliamo italiano.
Translation	I don't speak Spanish!
Task 1: score prediction	→ 25%
Task 2: error span prediction	→ I don't speak Spanish!
Task 3: post-editing	→ We don't speak Italian.

Table 1: Illustration of the three primary subtasks: segment-level quality score prediction, span-level error detection, and quality-informed post-editing. (The challenge sets subtask is not shown.)

The shared task consists of three primary subtasks shown in Table 1: (1) segment-level quality score prediction, (2) span-level translation error detection, and (3) quality-informed segment-level error correction. Curated evaluation data sets were provided for all three subtasks. These include test material obtained from the WMT25 General Machine Translation task as well as a collection of "challenge sets" that were developed by the organizers and members of the research community. A fourth subtask solicited the submission of these challenge sets.

The primary focus of this year's updated task

[1] www2.statmt.org/wmt25/mteval-subtask.html

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 414–461
November 8–9, 2025.

is on robust translation quality evaluation systems that can effectively detect translation errors generated by increasingly accurate LLM-based translators as well as handle previously unseen problems related to using LLMs for translation, such as output verbosity and incorrect output language. Newly, the evaluation data, originating from the General MT task, were intentionally chosen to be challenging for MT: they feature longer length, sourcing from some originally non-textual modalities, and translation into a wider variety of languages than in previous years.

Task 1 this year was designed to evaluate both MT metrics and QE systems. Reference-based automatic metrics score MT output by comparing the translation with a reference translation generated by human translators, who are instructed to translate "from scratch" without post-editing from MT. Reference-free quality estimation (QE) systems score translations solely based on their adherence to the source language. We collectively refer to all of these systems as "auto-raters" throughout the rest of the paper. All auto-raters were evaluated based on their agreement with human ratings when scoring MT systems and human translations at the system and segment level. In Task 2, systems are evaluated on their ability to detect and accurately annotate the spans of translation errors by contrasting them with the human error annotations. For Task 3, systems that correct a given translation are evaluated based on the quality of their post-edits, rewarding effective improvements with minimal changes.

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

- **Subtasks**: As illustrated in Table 1, we solicited participation in segment-level quality score prediction, span-level error detection, and quality-informed post-editing.

- **Language Pairs**: Based on the General MT shared task, this year covers 16 language pairs, many of which are novel: English→{Czech, Estonian, Icelandic, Egyptian Arabic, Bhojpuri, Maasai, Russian, Serbian Cyrillic, Ukrainian, Japanese, Chinese, Italian, Korean}, Czech→{German, Ukrainian}, and Japanese→Chinese. The domains are Literary (short story), News, Social, Speech (video transcripts), and Social.

- **Human Evaluation**: Human evaluation was done as part of the General MT task, and these same annotations were then reused for our shared task. For 14 of the language pairs, annotations were conducted with the ESA protocol (Kocmi et al., 2024); the remaining two language pairs were annotated using the MQM protocol (Freitag et al., 2021). ESA annotations included two sets of human annotations per translation.

Task	Level	Meta-metric
1	system	SPA
1	segment	acc^*_{eq}
2	char	F1
3	segment	Δ-COMET

Table 2: Each language was given equal weight in the overall average.

- **Meta-Evaluation**: Each of the three subtasks employs its own individual meta-evaluation methods, described in more detail in the respective sections later in the paper. Task 1 follows the same approach as last year's Metrics task and uses two primary measures: soft pairwise accuracy (SPA) at the system level (Thompson et al., 2024), and "group-by-item" segment-level accuracy with tie calibration (acc^*_{eq}) at the segment level (Deutsch et al., 2023). Task 2 follows a similar approach to last year's QE task, and calculates the character-level F1 score between the predicted errors and the gold error spans, weighted to allow for half points for correctly identified spans with incorrect error severity. Task 3 evaluates the quality of the correction of the original MT using ΔCOMET as the primary measure and Gain-to-Edit Ratio (GER) to quantify editing efficiency.

- **MTME**: Similar to last year, all the data for Tasks 1 and 2 has been uploaded to the `mt-metrics-eval` codebase (MTME),[2] and all results in this paper are calculated with this analysis tool. We encourage developers of auto-rater systems to use MTME for greater reproducibility.

Our main findings are:

- The prediction of quality scores, with or without references, remains a challenge. Strong LLM-based auto-raters now top the rankings at the system level, where the task is to accurately identify the better MT system. At

[2] github.com/google-research/mt-metrics-eval

the segment level, LLM-based auto-raters still underperform; this year, unlike in recent years, reference-based baseline metrics (YiSi-1, chrF, and BERTScore) fill out the top three rank clusters, outranking recently strong trained metrics. We provide some analysis of this surprising result, but further analysis is needed to develop a better understanding of this outcome (Section 4).

- Auto-raters continue to struggle with precise error detection, span annotation, and severity classification, with significant gaps with human performance and large variations across language pairs (Section 5). This underscores the complexity of the task and highlights the critical need for advances in automated error analysis that better align with human judgments.

- While automatic translation correction systems can improve translation quality, this improvement is often at the cost of diverging from human-generated reference translations, indicating a gap between automated systems and human lexical choices, and that improvement does not necessarily mean alignment with human preferences (Section 6).

- By using carefully-crafted challenge sets, it is shown that current automatic MT evaluation systems still exhibit major weaknesses, including susceptibility to fluent but semantically irrelevant content, systematic gender bias, instability on low-quality or corner-case outputs, and poor correlation with human judgments for low-resource languages (Section 7).

The rest of the paper is organized as follows. Section 2 introduces our subtasks. Section 3 presents the evaluation set and the MT systems whose output was judged. Following that, the baselines, participants, meta-evaluation procedure, and main results of each subtask are discussed in Section 4 for segment-level quality score prediction, in Section 5 for span-level error annotation, and in Section 6 for quality-informed post-editing. Section 7 presents the submitted challenge sets, and Section 8 concludes.

2 Tasks

The shared task this year consisted of three primary subtasks that address translation quality assessment from three perspectives: (1) segment-level quality score prediction, (2) span-level translation error detection, and (3) quality-informed segment-level error correction. An additional fourth subtask solicited the submission of challenge sets that identify where automated metrics and auto-rater systems fail. All subtasks are introduced below:

2.1 Segment-Level Quality Score Prediction

The goal of the segment-level quality prediction subtask is to predict a quality score for each source–target segment pair in the evaluation set, with a reference translation optionally being provided. Depending on the language pair, the participants were asked to predict either the Error Span Annotation (ESA) score (Kocmi et al., 2024) or the Multi-dimensional Quality Metrics (MQM) score (Freitag et al., 2021). Submissions are evaluated and ranked based on their prediction correlations with these human-annotated scores at both the segment and system levels.

2.2 Span-Level Error Detection

In this subtask, the goal is to predict the precise span of each translation error along with its severity. For this subtask we use the error spans obtained from the MQM and ESA human annotations generated for the General MT primary task as the target "gold standard". Participants were asked to predict both the error spans (start and end indices) as well as the error severities (major or minor) within each segment. Submissions are evaluated and ranked based on their ability to correctly identify the presence of errors, correctly mark the spans of any identified errors, and correctly identify the severity of each of these errors.

2.3 Quality-Informed Segment-level Error Correction

The overarching goal of this subtask is to correct the output of machine translation. Recent work shows that joint optimization for QE and APE helps improve the performance of both tasks (Deoghare et al., 2023, 2024). Furthermore, fine-grained QE signals can also be leveraged to apply limited corrections and help mitigate over-correction, known as a common problem in APE systems (Deoghare et al., 2025). We invited participants to submit systems capable of automatically generating corrections for machine-translated text, given the source, the MT-generated target, and a QE-generated quality annotation of the MT. The objective is to ex-

plore how quality information, including both error span annotations and Direct Assessment (DA) scores, can inform automated error correction. For instance, sentence-level quality scores may help identify which segments require correction, while span-level annotations can be used for fine-grained, pinpointed corrections. Participants were provided with the quality information. Submissions were evaluated and ranked based on the quality of the corrections they generate, with as few changes as possible.

2.4 Challenge Sets

For the fourth year, our shared task included a subtask involving challenge sets. 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. Whereas the standard evaluation of the shared task is conducted on test sets containing generic text from real-world content, the challenge set evaluation is based on test sets designed with the aim of revealing the abilities or the weaknesses of the metrics or evaluating particular translation phenomena. In order to shed light on different perspectives on evaluation, the subtask takes place in a decentralized manner: contrary to the main metric tasks, the test sets are not provided by the organizers but by different research teams, who are also responsible for analyzing and presenting the results (Section 7).

3 Evaluation Data

3.1 Data Sourcing and Translation

Similar to previous years' editions, the source sides, MT outputs, and reference texts for our shared task are mainly derived from the WMT25 General MT Shared Task (Kocmi et al., 2025a).

Newly this year, the source segments were automatically selected to be more challenging for translation systems, using a source-only difficulty estimator (Proietti et al., 2025). The test data domains cover news, literary (short stories), speech, and social. For the General MT shared task, some of this test material was multimodal: the speech data was provided as audio files with uncorrected automatic transcripts, and the social-media content was provided with screenshot images. However, for our shared task, we released only a text version of our evaluation set.

The selection of MT outputs was made based on an evaluation with automatic metrics (Kocmi et al., 2025b), giving slight prioritization to constrained (small, open-weight) systems. In keeping with our goal of exposing participants to a wide range of translation qualities and phenomena, we included output from around 20 different MT systems for each language pair. An exception is Task 3, which subsampled the original set for computational feasibility.

Reference translations were provided in 14 of our 16 language pairs, produced by professional translators from scratch. We ran English→Italian and English→Maasai as reference-free scenarios: the former because the General MT shared task intentionally did not produce any references, and the latter because the references produced by General MT were not yet available at the beginning of our evaluation period.

For more details regarding sourcing and translation of the test set, we refer the reader to the WMT25 General MT Shared Task (Kocmi et al., 2025a). All data has been released publicly.[3]

3.2 ESA and MQM Human Evaluation

This year, translations in most language pairs (English→{Czech, Estonian, Icelandic, Egyptian Arabic, Bhojpuri, Maasai, Russian, Serbian Cyrillic, Ukrainian, Japanese, Chinese, Italian} and Czech→{German, Ukrainian}) were evaluated with the ESA protocol (Kocmi et al., 2024). Japanese→Chinese and English→Korean were annotated with the MQM protocol (Freitag et al., 2021). The ESA protocol differs from MQM by not requiring the error categorization (fluency, punctuation, etc.) for each span and by providing a free-form 0% to 100% slider for assessing the final translation quality. Annotators were given guidance defining a score of 0% as "broken," 33% as "flawed," 66% as "good," and 100% as "perfect." In MQM, the translation score is instead derived mathematically from the count of error spans and their annotated severities. Generally, each minor error contributes –1 point while each major error counts as –5.

Our ESA annotations contain two judgments of each translation ("human1" and "human2"). This allows us to calculate intercoder agreement as a measure for task difficulty in each language and also provides a human "oracle" against which automated metrics can be compared (e.g. by treating

[3] github.com/wmt-conference/wmt25-general-mt

one human annotation as a "metric" and comparing it against actual auto-rater results). Additionally, the human annotations also contain control tasks (a fixed set of translations that all annotators annotated) designed to establish annotator reliability; this has been shown to work better than ad-hoc attention checks (Zouhar et al., 2025a).

In order to maximize differences between systems, human evaluation was limited to the top 50% of the most diversely translated inputs, computed with pairwise chrF between systems. Therefore, all source segments have either received two annotations for all selected systems, or none. This has the effect of avoiding spending human effort on annotating identical or similar translations, as well as translations that have little impact on the rankings of auto-rater systems (Zouhar et al., 2025b).

4 Task 1: Segment-Level Quality Score Prediction

This section presents the segment-level quality score prediction subtask in detail. We describe the auto-rater systems participating in the subtask in Section 4.1, our meta-evaluation procedure in Section 4.2, and our main results in Section 4.3. Section 4.4 reports some further analysis of the results beyond correlation and accuracy.

4.1 Participating Systems

We processed three distinct types of auto-rater systems participating in the segment-level quality score prediction subtask: baselines, official submissions, and "LLM as a judge" models. Each type is described in more detail below. A synthesized overview of all 48 systems is also given in Table 3, based on information provided by each participant at the time of submission. Full authoritative details are available in each team's separately prepared system description paper.

4.1.1 Baselines

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

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 [4] for system-level BLEU, and we used the `-sl` argument to obtain segment-level BLEU.
- **spBLEU (NLLB-Team et al., 2022)** is the BLEU score computed with subword tokenization by the standardized FLORES-200 SentencePiece models. We used the SacreBLEU command line to compute system-level spBLEU,[5] and we used the `-sl` argument to obtain segment-level spBLEU.
- **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[6] for system-level chrF and used the `-sl` argument to obtain segment-level chrF.

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.

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.

[4] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|v:2.3.1. 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.

[5] nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|v:2.3.1

[6] chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|sp.:no|v:2.3.1

Team	Auto-Rater Name	Purpose	Category	Backbone Model	Uses Ref?	Supervised?
Organizers	BERTSCORE	Baseline	embedding similarity	XLM-RoBERTa	Yes	No
Organizers	BLEU	Baseline	lexical overlap	—	Yes	No
Organizers	CHRF	Baseline	lexical overlap	—	Yes	No
Organizers	COMET22	Baseline	fine-tuned encoder	XLM-RoBERTa	Yes	Yes
Organizers	COMETKIWI22	Baseline	fine-tuned encoder	InfoXLM	No	Yes
Sentinel metrics	SENTINEL-CAND	Baseline	fine-tuned encoder	XLM-RoBERTa	No	Yes
Sentinel metrics	SENTINEL-SRC	Baseline	fine-tuned encoder	XLM-RoBERTa	No	Yes
Organizers	SPBLEU	Baseline	lexical overlap	—	Yes	No
Organizers	YISI-1	Baseline	embedding similarity	XLM-RoBERTa, BERT-zh	Yes	No
Phrase	ENSEMBLESLICK	Primary	fine-tuned LLM	GPT variants	No	Yes
Microsoft Translator	GEMBA-V2	Primary	LLM-based	GPT 4.1 mini	No	No
hw-tsc	HW-TSC	Primary	?	?	No	No
MetricX-25	METRICX-25	Primary	fine-tuned LLM	Gemma 3 12B	No?	Yes
CUNI	MR7.2.1	Primary	fine-tuned LLM	Gemma 3 27B IT	No	Yes
KIT-ETH-UMich	POLYCAND-2	Primary	fine-tuned encoder	XLM-RoBERTa	No	Yes
Sujal_and_Astha	RANKEDCOMET	Primary	fine-tuned encoder	XLM-RoBERTa	Yes?	Yes
DCU_ADAPT	ROBERTA-LS	Primary	fine-tuned encoder	XLM-RoBERTa	Yes	Yes
Nvidia-Nemo	SEGALE-QE	Primary	fine-tuned LLM	Gemma 3 12B	No	Yes
TASER	TASER-NO-REF	Primary	LLM-based	OpenAI o3	No	No
UvA-MT	UVA-MT	Primary	LLM-based	Gemma 3 12B	No	No?
Phrase	AUTOLQA	Secondary	fine-tuned LLM	GPT variants	No?	Yes
Sujal_and_Astha	BASECOMET	Secondary	fine-tuned encoder	XLM-RoBERTa	Yes?	Yes
CUNI	COLLABPLUS	Secondary	ensemble	—	No?	Yes
Phrase	COLLABSLICK	Secondary	fine-tuned LLM	GPT variants	No?	Yes
hw-tsc	HW-TSC-BASE	Secondary	?	?	No?	No
hw-tsc	HW-TSC-MAX	Secondary	?	?	No?	No
DCU_ADAPT	LONG-CONTEXT	Secondary	fine-tuned	?	Yes?	Yes
MetricX-25	METRICX-25-QE	Secondary	fine-tuned LLM	Gemma 3 12B	No	Yes
MetricX-25	METRICX-25-REF	Secondary	fine-tuned LLM	Gemma 3 12B	Yes	Yes
CUNI	MR6	Secondary	fine-tuned LLM	Gemma 3 27B IT	No?	Yes
KIT-ETH-UMich	POLYCAND-1	Secondary	fine-tuned encoder	XLM-RoBERTa	No	Yes
KIT-ETH-UMich	POLYIC-3	Secondary	fine-tuned encoder	XLM-RoBERTa	No	Yes
Nvidia-Nemo	Q_MQM	Secondary	LLM-based	Qwen 3	No?	No
Nvidia-Nemo	Q_RELATIVE-MQM	Secondary	LLM-based	Qwen 3	No	No
DCU_ADAPT	ROBERTA-MULTI	Secondary	fine-tuned encoder	XLM-RoBERTa	Yes?	Yes
TASER	TASER-REF	Secondary	LLM-based	OpenAI o3	Yes?	No
Organizers	AYAEXPANSE-32B	LLM	LLM	AyaExpanse 32B	No	No
Organizers	AYAEXPANSE-8B	LLM	LLM	AyaExpanse 8B	No	No
Organizers	CLAUDE-4	LLM	LLM	Claude 4	No	No
Organizers	COMMANDA	LLM	LLM	CommandA	No	No
Organizers	COMMANDR7B	LLM	LLM	CommandR 7B	No	No
Organizers	DEEPSEEK-V3	LLM	LLM	DeepSeek V3	No	No
Organizers	GPT-4.1	LLM	LLM	GPT 4.1	No	No
Organizers	LLAMA-3.1-8B	LLM	LLM	Llama 3.1 8B	No	No
Organizers	LLAMA-4-MAVERICK	LLM	LLM	Llama 4 Maverick	No	No
Organizers	MISTRAL-7B	LLM	LLM	Mistral 7B	No	No
Organizers	QWEN2.5-7B	LLM	LLM	Qwen 2.5 7B	No	No
Organizers	QWEN3-235B	LLM	LLM	Qwen 3 235B	No	No

Table 3: Summary of Task 1 participants. We distinguish four different purposes of participation: as a baseline, as an official primary submission, as an official secondary submission, or as an "LLM as a judge." Basic self-submitted properties of each entrant are summarized above, sorted by auto-rater name.

Sentinel baselines We also included two metrics from the Sentinel family. Unlike the other baselines, Sentinel metrics are intentionally formulated to lack important information when assigning their scores. They are instead meant as a probing mechanism to highlight evaluation scenarios that may be "too easy" or that are prone to spurious correlations, if the Sentinel metrics place competitively among other evaluators that have access to more complete information.

- **SENTINEL-SRC-25 (Proietti et al., 2025)** predicts the quality of a translation solely based on its source string, without considering the reference or even the translation itself. It is an updated version of the original SENTINEL-SRC: a regression model based on XLM-RoBERTa, trained with data from previous WMT editions up through and including the WMT 2024 test set.

- **SENTINEL-CAND (Perrella et al., 2024)** assesses the quality of a translation based on the output string alone, without taking the source or reference into account. It is also based on XLM-RoBERTa, trained with WMT data up

through 2022.

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).

4.1.2 Official Submissions

Each team participating in Task 1 was allowed to submit one primary and up to two secondary systems for meta-evaluation. The primary systems are described below. Secondary systems are included in the general tabular overview (Table 3).

ENSEMBLESLICK (Hrabal et al., 2025) For Task 1, this system uses a combination of Phrase proprietary fine-tuned GTE and similar models and fine-tuned GPT-4o-mini.

GEMBA-V2 (Junczys-Dowmunt, 2025) GEMBA-V2 is an updated version of GEMBA (Kocmi and Federmann, 2023).

HW-TSC (Luo et al., 2025) This system's approach integrates sentence segmentation tools and dynamic programming to construct sentence-level alignments between source and translated texts, then adapts sentence-level evaluation models to document-level assessment via sliding-window aggregation.

METRICX-25 (Juraska et al., 2025) METRICX-25 is an encoder-only regression model initialized from Gemma 3 12B and fine-tuned on publicly available DA and MQM scores from WMT 2015–23 in a two-stage fashion. Similar to METRICX-24, the first stage uses z-normalized DA scores, and the second stage uses a mixture of raw DA scores (rescaled to the MQM range of 0–25) and MQM scores. Due to the dual nature of meta-evaluation this year (ESA/DA vs. MQM), a score type indication is included in the input, indicating for each training example whether it corresponds to a DA or MQM score.

MR7.2.1 (Hrabal et al., 2025) This submission experimented with the Gemma 3 27B IT model prompted using the DSPy framework and using its MIPROv2 optimizer. The system first generates seven 0–10 integer scores for various aspects of the translation (e.g. "accuracy and completeness" or "fluency and coherence"). Afterwards, it generates the overall 0–100 score.

POLYCAND-2 (Züfle et al., 2025) The supervised reference-less metric $\text{COMET}_{\text{poly-}*}$ has similar architecture and training to standard COMET but incorporates additional information beyond one single translation. $\text{COMET}_{\text{poly-cand2}}$ incorporates two alternative translations of the same source segment (provided by other translation systems) to better contextualize and assess the quality of the translation being scored. The metric was trained on a limited combination of DA, ESA, and MQM data on a unified scale.

RANKEDCOMET (Maharjan and Shrestha, 2025) This system is based on the pre-trained Unbabel/wmt22-comet-da model, deployed in a zero-shot inference setting. Raw segment-level quality scores are generated and then post-processed with per-language-pair rank normalization. This method transforms raw scores into a calibrated distribution that significantly improved correlation with the preliminary evaluation metrics.

ROBERTA-LS (Haq and Osuji, 2025) ROBERTA-LS (Roberta Long-Span) is a reference-based evaluation metric built using the COMET framework. Designed to provide multi-sentence quality scores, it is trained on augmented long-context data that captures translation quality beyond isolated sentences. To construct the long-span MT evaluation dataset, adjacent short segments are concatenated, and a multi-segment quality score is computed as a length-weighted average of their original scores. Unbabel/wmt22-comet-da and XLM-RoBERTa-base are fine-tuned on the augmented data.

SEGALE-QE (Yan et al., 2025) This system extends METRICX to long texts by adding a pipeline before running the metric. It first segments the data down to individual sentences with Ersatz, then runs Vecalign to align system translations to the source. Vecalign's deletion penalty is adaptively adjusted to obtain good alignments that exclude over/under-translation to the maximum extent possible. When over/under-translated sentences are identified, they are assigned a score of 25. Individual sentence scores are then averaged to form the score for the long-form translation pair.

TASER-NO-REF (Maheswaran et al., 2025) TASER (Translation Assessment via Systematic Evaluation and Reasoning) is a Large Reasoning Model–based metric for translation quality assessment. This metric uses OpenAI's o3 to estimate the quality of a translation in reference-free scenarios. It posits that LRMs are capable of better assessing the quality of translations than vanilla LLMs with advanced prompting strategies.

UVA-MT (Wu and Monz, 2025) This system calibrates quality estimation and likelihood on the google/gemma3-12b-it model, then directly uses the token average likelihood as a metric for quality estimation. No human annotation data is used; the only reliance is on a translation's likelihood as the metric. The same resulting model was also submitted to the WMT 2025 General Translation Task, meaning that it grades its own output as part of the segment-level quality prediction task.

4.1.3 LLMs as Judges

As a third category of system, the shared task organizers obtained quality scores on our test set from 12 different publicly available large language models, using their standard APIs and a templated prompt. These submissions test the ability of general-purpose LLMs as judges of translation quality without fine-tuning or few-shot examples.

LLMs for which we obtained quality scores are: AyaExpanse 8B, AyaExpanse 32B, Claude 4, Command A, Command R7B, DeepSeek V3, GPT 4.1, Llama 3.1 8B, Llama 4 Maverick, Mistral 7B, Qwen 2.5 7B, and Qwen 3 235B. The templated ESA-like prompt is given in Appendix A.

4.2 Meta-Evaluation

The goal of auto-rater meta-evaluation is to quantify how well automatic systems agree with human ratings of translation quality. There are a multitude of ways to approach this problem, as evidenced by the variety of solutions proposed by previous years' editions of the shared task.[7] Ranking-based approaches (traditionally Spearman's ρ, Kendall's τ, or pairwise accuracy) assume the least about the relative shapes of the score distributions: only the directionality matters, and the magnitude of difference is ignored. Linear correlation (traditionally Pearson's r) captures magnitude but thereby assumes a constant slope to the scores and can be unduly influenced by outliers (Mathur et al., 2020).

We follow the same approach as last year to this year's meta-evaluation of Task 1, focusing on improved ranking-based methods.

At the system level, we use soft pairwise accuracy, or SPA (Thompson et al., 2024). SPA uses p-values as a proxy for certainty about the difference between two systems, calculated over both the auto-rater and human scores. This rewards auto-raters that result in the same statistical conclusion as the human scores. However, computation of the p-values requires repeated resampling of segments in order to determine the statistical range of system-level scores. For efficiency of meta-evaluation, SPA *averages* the segment-level scores in each resample as a proxy for the system-level score. Such averaging is a technically incorrect aggregation method for BLEU, CHRF, and a number of other submissions that self-reported that they employ some more complicated methodology.

At the segment level, we again follow last year's process and meta-evaluate metrics using "group-by-item" segment-level accuracy with tie calibration (Deutsch et al., 2023), denoted acc^*_{eq}. Group-by-item processing, recommended by Perrella et al. (2024) avoids pairwise comparisons between translations originating from different source segments.

Because SPA and acc^*_{eq}meta-metrics are based on ranking, we do not perform any normalization on the raw scores output by the auto-raters or annotated by the humans.

We assign ranks to auto-raters based on their significance clusters in the same way that we did last year. We compare all pairs of auto-raters and determine whether the difference in their correlation scores is significant according to the PERM-BOTH hypothesis test of Deutsch et al. (2021). We use 1000 re-sampling runs and set $p = 0.05$. As advocated by Wei et al. (2022), we divide the sample into blocks of 100, compute significance after each block (cumulative over all blocks sampled so far), and stop early if the p-value is < 0.02 or > 0.50. To calculate p-values for SPA, we use a paired permutation test (Noreen, 1989) with 1000 resamples.

Given the significance results (p-values) for all pairs of auto-raters, ranks are assigned starting with the highest-scoring auto-rater. We move down the list of auto-raters in descending order by score, assigning rank 1 to all auto-raters until we encounter the first one that is significantly different from any that have been visited so far. That auto-rater is

[7] See Section 5 of Thompson et al. (2024) and Table 1 of Deutsch et al. (2023) for nice summaries of the approaches taken in prior WMT shared tasks to meta-evaluation at, respectively, the system and segment level.

assigned rank 2, and the process is repeated. This continues until all auto-raters have been assigned a rank. Note that this is a greedy algorithm, and hence it can place two auto-raters that are statistically indistinguishable in different clusters.

The code for running the meta-evaluation is available in the `mt-metrics-eval` library.[8]

While the segments in language pairs evaluated with ESA received two independent human judgments (as per Section 3.2), most of the results and analyses presented in Section 4.3 and Section 4.4 are based on the first complete annotation that we received, which we refer to as "human1." The second set of human judgements ("human2") did not arrive in time to permit a complete analysis with the `mt-metrics-eval` package.

4.3 Main Results

Summarized results for the quality score prediction subtask are shown in Table 4 and Table 5. Table 4 reports average performance across the 14 language pairs for which references were provided, while Table 5 covers the remaining two reference-less language pairs. (Since we do not have complete information about which auto-raters make use of the reference or may do so optionally, we list in Table 5 all the entrants that submitted scores for reference-free language pairs.) Note that three participating systems (ROBERTA-LS, LONG-CONTEXT, and ROBERTA-MULTI) returned output for only three language pairs and are thus not included in either summary table for lack of a fair comparison. Full detailed results broken down by individual language pair are given in Table 19 (part 1) and Table 20 (part 2) in Appendix B; all participating systems appear there.[9]

A striking pattern in this year's results is the strong performance of many baseline systems — especially those based on lexical overlap or embedding similarity. YISI-1, CHRF, and BERTSCORE fill out the top three rank clusters when judged at the segment level in the presence of references. General-purpose LLMs do quite poorly in terms of correlation with human judgments at the segment level, but their system-level performance is better, led by GPT 4.1 and Claude 4. GPT 4.1 and TASER-REF, a reasoning-based model, achieve top-tier performance on system-level correlation when a reference is present; reference-free models are led by GEMBA-V2.[10] Sentinel models, as desired, rank lowly throughout.

	Avg All		Avg Sys		Avg Seg	
	Rank	Corr	Rank	Corr	Rank	Corr
Baselines						
YiSi-1	2	0.674	4	0.791	1	0.558
chrF	2	0.672	4	0.789	2	0.554
spBLEU	3	0.668	5	0.784	4	0.551
BERTScore	4	0.662	6	0.770	3	0.553
BLEU	5	0.657	6	0.770	6	0.543
COMET22	8	0.624	9	0.709	8	0.539
sentinel-cand	17	0.533	16	0.572	17	0.494
COMETKiwi22	19	0.505	18	0.526	20	0.484
sentinel-src	25	0.351	19	0.509	37	0.193
Primary						
GEMBA-v2	2	0.672	3	0.811	9	0.533
TASER-No-Ref	3	0.666	2	0.833	16	0.499
rankedCOMET	6	0.627	8	0.716	8	0.539
MetricX-25	8	0.621	9	0.711	10	0.530
mr7_2_1	9	0.614	6	0.760	24	0.467
SEGALE-QE	13	0.581	12	0.654	12	0.509
Polycand-2	14	0.566	13	0.626	13	0.506
Q_Relative-MQM	15	0.564	7	0.737	28	0.391
EnsembleSlick	17	0.539	15	0.600	23	0.478
hw-tsc	18	0.524	17	0.557	18	0.490
UvA-MT	21	0.465	20	0.466	25	0.464
Secondary						
TASER-Ref	1	0.698	1	0.846	5	0.549
MetricX-25-Ref	6	0.633	8	0.727	7	0.539
baseCOMET	7	0.624	10	0.709	7	0.539
MetricX-25-QE	10	0.602	11	0.681	11	0.524
mr6	10	0.598	7	0.738	26	0.458
Q_MQM	14	0.568	7	0.736	27	0.399
Polyic-3	16	0.555	14	0.607	14	0.503
AutoLQA	16	0.553	10	0.707	27	0.398
Polycand-1	16	0.554	14	0.606	15	0.501
CollabPlus	16	0.548	13	0.612	20	0.485
CollabSlick	16	0.548	14	0.609	19	0.487
hw-tsc-max	19	0.509	18	0.536	21	0.483
hw-tsc-base	20	0.499	19	0.518	22	0.479
LLM-as-a-judge						
GPT-4_1	3	0.669	1	0.849	18	0.489
CommandA	11	0.597	3	0.812	29	0.382
Claude-4	12	0.593	2	0.833	31	0.352
DeepSeek-V3	13	0.582	4	0.797	30	0.368
Qwen3-235B	13	0.579	4	0.790	30	0.368
Qwen2_5-7B	19	0.507	12	0.667	32	0.347
AyaExpanse-32B	19	0.500	7	0.732	34	0.269
Llama-3_1-8B	21	0.466	12	0.663	34	0.269
Llama-4-Maverick	22	0.453	7	0.730	38	0.176
CommandR7B	23	0.408	16	0.568	35	0.248
Mistral-7B	23	0.401	18	0.527	33	0.274
AyaExpanse-8B	24	0.387	15	0.576	36	0.199

Table 4: Task 1 results summary against the "human1" annotation for all language pairs with references.

[8] github.com/google-research/mt-metrics-eval

[9] We also show in Appendix B the results on the ESA language pairs using the "human2" annotation as the gold standard.

[10] TASER-REF also ranks first or second in reference-free evaluation; even though it is labeled as a reference-using metric, it was submitted with scores for the segments without references as well.

	Avg All		Avg Sys		Avg Seg	
	Rank	Corr	Rank	Corr	Rank	Corr
Baselines						
sentinel-cand	8	0.542	6	0.593	7	0.492
COMETKiwi22	10	0.501	9	0.517	8	0.485
sentinel-src	12	0.417	9	0.501	23	0.333
Primary						
GEMBA-v2	1	0.638	1	0.764	2	0.512
TASER-No-Ref	3	0.601	3	0.710	7	0.493
mr7_2_1	4	0.581	3	0.702	11	0.460
SEGALE-QE	4	0.570	5	0.632	3	0.508
EnsembleSlick	7	0.550	5	0.609	7	0.491
Q_Relative-MQM	7	0.549	4	0.686	18	0.413
MetricX-25	7	0.548	7	0.583	2	0.514
Polycand-2	7	0.547	6	0.599	5	0.495
rankedCOMET	8	0.542	7	0.592	7	0.493
hw-tsc	8	0.538	7	0.584	7	0.491
UvA-MT	11	0.485	10	0.494	10	0.476
Secondary						
TASER-Ref	1	0.633	2	0.738	1	0.528
Q_MQM	4	0.578	2	0.740	17	0.415
mr6	5	0.569	4	0.682	12	0.456
MetricX-25-QE	6	0.554	6	0.594	2	0.514
CollabSlick	6	0.553	5	0.609	4	0.498
baseCOMET	7	0.549	6	0.605	6	0.493
Polycand-1	8	0.536	7	0.579	7	0.493
Polyic-3	9	0.532	8	0.570	5	0.495
CollabPlus	9	0.528	8	0.557	4	0.498
AutoLQA	9	0.525	6	0.597	12	0.453
hw-tsc-max	10	0.512	9	0.541	9	0.483
hw-tsc-base	10	0.512	9	0.541	9	0.483
LLM-as-a-judge						
GPT-4_1	2	0.611	2	0.725	4	0.496
CommandA	4	0.577	3	0.711	13	0.443
Claude-4	4	0.574	2	0.729	16	0.419
Qwen3-235B	4	0.573	3	0.712	14	0.434
DeepSeek-V3	5	0.565	3	0.716	17	0.413
Qwen2_5-7B	5	0.562	3	0.696	15	0.428
AyaExpanse-32B	7	0.543	3	0.702	19	0.383
CommandR7B	9	0.529	4	0.685	20	0.373
Llama-3_1-8B	10	0.513	4	0.662	21	0.364
Llama-4-Maverick	11	0.485	4	0.669	24	0.301
Mistral-7B	11	0.484	6	0.594	20	0.373
AyaExpanse-8B	11	0.473	5	0.609	22	0.337

Table 5: Task 1 results summary against the "human1" annotations for all language pairs without references (i.e. English–Italian and English–Maasai).

Divergence in these results compared to recent years may be due to a variety of causes. In particular, we note that the source texts were intentionally chosen to be difficult, that they consist of longer paragraph-like segments, and that there are therefore fewer segments available for scoring in each language pair than in the past. Further, we ran the quality score prediction task in a wider variety of language pairs, with a correspondingly wider variety in the quality of the MT output being judged.

We will further explore a few interesting facets of the results in the following section.

4.4 Analysis

In this section, we discuss the performance of MT auto-rater systems from several additional perspectives, in order to interpret our results, to provide further insights on strength and weakness of various classes of auto-raters, and to shed light on upcoming challenges in automated translation quality evaluation research.

4.4.1 SPA vs. Pairwise Accuracy

Because the results in Section 4.3 differ from those obtained in other recent years — notably on the relative strength of string-based metrics — we ran a contrastive evaluation where the system-level meta-metric was changed from SPA to "hard" pairwise accuracy instead. System-level pairwise accuracy (Kocmi et al., 2021) was used as a meta-evaluation metric in the WMT metrics task from 2021 through 2023. This method, similarly to SPA, compares MT system pair ranking decisions between humans and auto-raters, but it ignores the magnitude of the human and auto-rater score differences. Pairwise accuracy also does not require any resampling or averaging of segment-level scores to create proxies for system-level scores.

Table 6 shows the correlation results obtained by using each system-level meta-metric, for language pairs that were provided with references, against the "human1" annotations as the gold standard. The "SPA" columns of the table are equivalent to the system-level data shown in Table 4, repeated here for easy side-by-side comparison. The "Pair Acc" columns show the analogous results using pairwise accuracy. (Note that the use of pairwise accuracy leads to a smaller number of statistically significant auto-rater clusters, so the rank ordinals are not comparable between the "SPA" and "Pair Acc" columns.) The right-most "Diff" column shows the differences in correlation between the two settings.

SPA and pairwise accuracy produced very similar correlation scores and overall rankings of auto-raters. Our contrastive experiments therefore did not shed light on why string-based metrics are performing better than expected. The main difference we observe is that SPA produced substantially more statistically significant comparisons, resulting in twice as many (20 vs. 10) significance clusters. This finding is consistent with Thompson et al. (2024).

	SPA		Pair Acc		Diff
	Rank	Corr	Rank	Corr	ΔCorr
Baselines					
YiSi-1	4	0.791	3	0.795	0.004
chrF	4	0.789	3	0.783	−0.006
spBLEU	5	0.784	3	0.780	−0.004
BERTScore	6	0.770	3	0.772	0.002
BLEU	6	0.770	3	0.767	−0.003
COMET22	9	0.709	5	0.701	−0.008
sentinel-cand	16	0.572	8	0.573	0.001
COMETKiwi22	18	0.526	9	0.515	−0.011
sentinel-src	19	0.509	9	0.476	−0.033
Primary					
TASER-No-Ref	2	0.833	2	0.828	−0.005
GEMBA-v2	3	0.811	2	0.815	0.004
mr7_2_1	6	0.760	4	0.760	0.000
MetricX-25	9	0.711	5	0.705	−0.006
rankedCOMET	8	0.716	5	0.703	−0.013
SEGALE-QE	12	0.654	6	0.654	0.000
Polycand-2	13	0.626	6	0.627	0.001
EnsembleSlick	15	0.600	8	0.597	−0.003
hw-tsc	17	0.557	8	0.555	−0.002
UvA-MT	20	0.466	10	0.468	0.002
Secondary					
TASER-Ref	1	0.846	1	0.846	0.000
Q_MQM	7	0.736	4	0.742	0.006
mr6	7	0.738	4	0.741	0.003
Q_Relative-MQM	7	0.737	4	0.740	0.003
MetricX-25-Ref	8	0.727	4	0.720	−0.007
AutoLQA	10	0.707	5	0.708	0.001
baseCOMET	10	0.709	5	0.702	−0.007
MetricX-25-QE	11	0.681	6	0.673	−0.008
CollabSlick	14	0.609	7	0.614	0.005
CollabPlus	13	0.612	7	0.613	0.001
Polycand-1	14	0.606	7	0.608	0.002
Polyic-3	14	0.607	7	0.604	−0.003
hw-tsc-max	18	0.536	9	0.538	0.002
hw-tsc-base	19	0.518	9	0.524	0.006
LLM-as-a-judge					
GPT-4_1	1	0.849	1	0.849	0.000
Claude-4	2	0.833	1	0.839	0.006
CommandA	3	0.812	2	0.813	0.001
DeepSeek-V3	4	0.797	3	0.802	0.005
Qwen3-235B	4	0.790	3	0.794	0.004
Llama-4-Maverick	7	0.730	4	0.741	0.011
AyaExpanse-32B	7	0.732	4	0.734	0.002
Qwen2_5-7B	12	0.667	6	0.669	0.002
Llama-3_1-8B	12	0.663	6	0.660	−0.003
CommandR7B	16	0.568	8	0.575	0.007
AyaExpanse-8B	15	0.576	8	0.550	−0.026
Mistral-7B	18	0.527	9	0.507	−0.020

Table 6: System-level correlation using either SPA (equivalent to Table 4) or pairwise accuracy, against the "human1" annotation for all language pairs with references. The scores and overall rankings are quite similar for SPA and pairwise accuracy, but SPA produces substantially more (20 vs. 10) significance clusters.

	All MT		Select MT		Diff
	Rank	Corr	Rank	Corr	ΔCorr
Baselines					
spBLEU	5	0.784	3	0.789	0.006
YiSi-1	4	0.791	3	0.789	−0.002
chrF	4	0.789	3	0.786	−0.003
COMET22	9	0.709	4	0.770	0.060
BLEU	6	0.770	4	0.769	−0.001
BERTScore	6	0.770	4	0.764	−0.006
sentinel-cand	16	0.572	7	0.684	0.112
COMETKiwi22	18	0.526	10	0.551	0.025
sentinel-src	19	0.509	11	0.495	−0.014
Primary					
TASER-No-Ref	2	0.833	1	0.836	0.003
GEMBA-v2	3	0.811	2	0.819	0.008
rankedCOMET	8	0.716	4	0.775	0.059
mr7_2_1	6	0.760	4	0.761	0.001
MetricX-25	9	0.711	5	0.756	0.045
Q_Relative-MQM	7	0.737	6	0.719	−0.017
SEGALE-QE	12	0.654	6	0.693	0.039
Polycand-2	13	0.626	7	0.687	0.062
EnsembleSlick	15	0.600	7	0.669	0.069
hw-tsc	17	0.557	8	0.652	0.095
UvA-MT	20	0.466	12	0.374	−0.091
Secondary					
TASER-Ref	1	0.846	1	0.840	−0.006
baseCOMET	10	0.709	4	0.770	0.061
MetricX-25-Ref	8	0.727	5	0.749	0.022
MetricX-25-QE	11	0.681	5	0.747	0.066
AutoLQA	10	0.707	5	0.743	0.036
mr6	7	0.738	5	0.740	0.002
Q_MQM	7	0.736	6	0.719	−0.017
Polyic-3	14	0.607	6	0.699	0.092
Polycand-1	14	0.606	7	0.684	0.078
CollabPlus	13	0.612	7	0.676	0.064
CollabSlick	14	0.609	7	0.671	0.061
hw-tsc-max	18	0.536	9	0.602	0.067
hw-tsc-base	19	0.518	10	0.573	0.055
LLM-as-a-judge					
GPT-4_1	1	0.849	1	0.840	−0.009
Claude-4	2	0.833	1	0.834	0.002
CommandA	3	0.812	2	0.820	0.008
DeepSeek-V3	4	0.797	2	0.814	0.017
Qwen3-235B	4	0.790	3	0.796	0.006
AyaExpanse-32B	7	0.732	5	0.750	0.019
Llama-4-Maverick	7	0.730	6	0.723	−0.002
Llama-3_1-8B	12	0.663	8	0.633	−0.030
Qwen2_5-7B	12	0.667	8	0.631	−0.035
AyaExpanse-8B	15	0.576	10	0.530	−0.045
Mistral-7B	18	0.527	10	0.524	−0.004
CommandR7B	16	0.568	11	0.478	−0.090

Table 7: Average system-level correlations using either all available MT output (equal to Table 4) or only output from selected MT systems unlikely to have hill-climbed on metrics. Correlations are computed against "human1" annotations.

4.4.2 MT Systems That Hill-Climb Metrics

Some systems in the WMT General MT task (Kocmi et al., 2025a), whose output we rely on to create this task's test set, are tuned against automatic metrics (for example, as part of the reward model). This presents a challenge when the same automatic metric is now asked to judge the qual-

ity of the MT: a bias towards the translations that have been specially optimized towards it could negatively affect the metric's correlation with independent human judgments.

To investigate this effect, we conducted a follow-up analysis in which we calculated auto-rater performance only when judging output from a population of MT systems that are highly unlikely to be metric-tuned. These selected MT systems primarily consist of general-purpose LLMs and publicly available MT services. We repeated the system-level meta-evaluation from Section 4.2, using SPA, on this reduced portion of our test set.

Table 7 shows a comparison of the results in the two cases. The "All MT" column represents the rank and average system-level correlation of each participant for the 14 language pairs in the original test set that provide reference translations. (This section of the table is equivalent to the system-level columns of Table 4.) In the "Select MT" column, we display the analogous results on the smaller test set. (Note that the smaller test set leads to a smaller number of statistically significant metric clusters, so the rank ordinals are not comparable between the two columns.) The right-most "Diff" column shows the differences in average correlation between the two settings.

As expected, metrics that are the most likely targets of MT hill-climbing see their correlations with human judgments improve once we remove the affected MT systems. This is true most visibly for the numerous variations of COMET — including all three POLYCAND* auto-raters, all of which rank in the top 10 "most improved." Conversely, the smallest changes in average correlation tend to come from classic string- or embedding-based metrics, which are unlikely to serve as modern-day MT optimization targets, as well as TASER variants and high-performing general-purpose LLMs.

With these results in mind, we caution MT practitioners against evaluating system variants according to the same metrics that played any role in the systems' training process.

4.4.3 Detecting Catastrophic Translations

The distribution of translation quality varies greatly across languages. For example, high-resource languages in our test set tend to come with the most translations that are near perfect, while even state-of-the-art MT systems struggle with lower-resource languages or languages and domains not previously in WMT. We show this distribution of human-annotated ESA scores in Figure 1. This is a problem for trained auto-rater systems, which underperform in unseen domains and tend to follow the language distribution; they are thus likely to score translations into a low-resource language as lower in quality and they have greater variance (Zouhar et al., 2024a,b).

This year's language pairs created a new issue for MT systems: incorrect output language. Specifically, for some language pairs, some MT systems outputted the wrong language, dialect, or script.

Auto-Rater	en-ar	en-bho	en-sr	en-et	en-is	en-ru
Human	98%	78%	86%	24%	46%	13%
Claude-4	77%	61%	73%	19%	49%	16%
GPT-4	84%	39%	74%	23%	54%	15%
TASER-Ref	77%	50%	56%	21%	53%	16%
COMETKiwi22	84%	58%	55%	14%	41%	13%
Polyic-3	81%	67%	33%	15%	48%	12%
UvA-MT	76%	49%	69%	16%	36%	10%
Polycand-2	80%	48%	48%	16%	48%	12%
MetricX-25-Ref	79%	47%	51%	17%	44%	13%
DeepSeek-V3	76%	28%	69%	16%	45%	14%
MetricX-25-QE	77%	45%	51%	17%	45%	13%
BERTScore	84%	51%	50%	16%	36%	10%
Polycand-1	80%	53%	39%	14%	48%	12%
hw-tsc	78%	56%	48%	13%	37%	12%
SEGALE-QE	76%	49%	43%	17%	46%	12%
hw-tsc-base	79%	56%	46%	13%	35%	12%
YiSi-1	79%	43%	47%	18%	43%	11%
hw-tsc-max	77%	56%	46%	14%	35%	12%
CommandA	77%	24%	67%	16%	38%	14%
MetricX-25	77%	37%	49%	15%	42%	13%
sentinel-cand	80%	54%	32%	13%	41%	11%
COMET22	79%	30%	42%	18%	44%	12%
rankedCOMET	79%	30%	42%	18%	44%	12%
baseCOMET	79%	30%	42%	18%	44%	12%
mr7_2_1	78%	23%	53%	19%	37%	15%
Llama-4-Maverick	76%	25%	53%	13%	43%	13%
mr6	76%	27%	52%	16%	36%	13%
Qwen3-235B	76%	22%	50%	19%	35%	14%
Q_Relative-MQM	76%	23%	56%	12%	35%	14%
Q_MQM	76%	23%	45%	11%	35%	15%
GEMBA-v2	76%	25%	29%	13%	47%	14%
chrF	80%	36%	17%	17%	41%	9%
TASER-No-Ref	76%	36%	23%	9%	45%	12%
spBLEU	82%	34%	18%	18%	38%	10%
CollabSlick	76%	52%	8%	16%	35%	10%
CollabPlus	76%	51%	8%	14%	36%	11%
Qwen2	76%	32%	35%	14%	28%	11%
BLEU	80%	39%	18%	14%	35%	9%
EnsembleSlick	76%	50%	7%	16%	35%	10%
Mistral-7B	77%	31%	34%	9%	26%	13%
AyaExpanse-32B	76%	19%	40%	10%	29%	13%
Llama-3	76%	19%	29%	15%	31%	11%
AyaExpanse-8B	76%	21%	27%	11%	28%	13%
CommandR7B	77%	31%	24%	8%	27%	11%
AutoLQA	76%	19%	17%	11%	34%	11%
sentinel-src	76%	18%	8%	14%	26%	13%

Table 8: Ability of auto-raters to detect catastrophic translations (best threshold for F_1). Rows are ordered by average performance; auto-raters perform worse than human for languages with bimodal distributions (Figure 1).

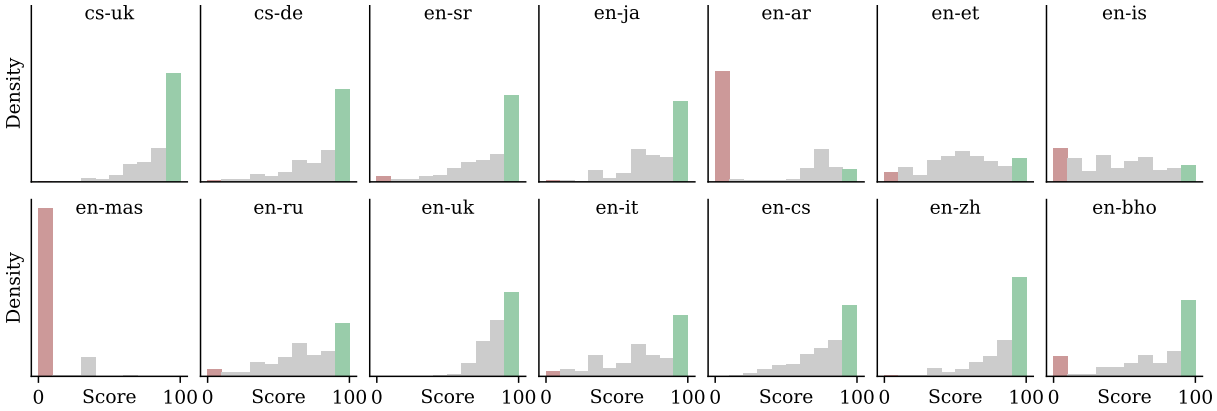

Figure 1: Human score distribution across languages evaluated with the ESA protocol. The pink bar corresponds to translation with less than 10 ESA points, and the green bar shows translations with above 90 ESA points.

Examples include English→Egyptian Arabic (oftentimes incorrectly translated as modern standard Arabic), English→Bhojpuri (incorrectly translated as a mixture of languages), or English→Serbian (incorrectly translated into a Latin script). These systems still made it into the human evaluation because the automatic evaluation filter used did not flag any major issues. In this section, we discuss the failure of auto-rater systems to detect catastrophic translation failure, in use-cases similar to those used by the General MT shared task.

We investigate select languages that do have a notable proportion of catastrophic translations, defined as those receiving an ESA score lower than 10. For those, we collect the set of catastrophic translations and select the threshold for each auto-rater to classify a catastrophic translation that maximizes the F_1 score (true positive = ESA score < 10). This threshold would, in theory, be possible to be used for identifying catastrophic translations without any human input. The results are shown in Table 8. For languages where the mismatch in languages, dialects, and scripts was a critical issue (Arabic, Bhojpuri, Serbian), the second human annotator was able to detect the catastrophic translations much better than automatic systems. This leaves headroom for improvements for auto-rater systems, which have to be able to score translations from state-of-the-art LLMs that might not always follow the instructions.

4.4.4 Utility of the Reference

Having investigated auto-rater systems capabilities in the face of especially poor MT output, we now turn to the complementary question: a study of our submissions' abilities to cope with poor *reference* translations.

Auto-Rater	**ΔCorr**	**Auto-Rater**	**ΔCorr**
YiSi-1	−0.163	CommandR7B	0.201
chrF	−0.152	Llama-3_1-8B	0.199
BLEU	−0.151	AyaExpanse-8B	0.189
BERTScore	−0.148	Qwen2_5-7B	0.178
spBLEU	−0.143	Mistral-7B	0.164
baseCOMET	−0.043	mr7_2_1	0.160
COMET22	−0.043	Q_MQM	0.144
hw-tsc-max	−0.036	Q_Relative-MQM	0.139
GPT-4_1	−0.030	AyaExpanse-32B	0.137
rankedCOMET	−0.030	mr6	0.112

Table 9: Auto-raters recording the largest drops (left) and gains (right) in average system-level correlation between language pairs with high- and low-quality references.

We would expect to see a divergence in performance between auto-raters that adhere closely to the reference and those that are reference-free. When the provided reference is itself of poor quality, auto-raters in the first group may be misled into misjudging the MT output. When the reference is quite accurate, on the other hand, auto-raters in the second group may suffer from not being able to consult it. Below we examine each of these cases individually.

For this analysis, we divide the language pairs in our test set into groups based on the *references'* performance in the human evaluation. In the WMT General Translation task (Kocmi et al., 2025a), the reference translation was judged to fall alone into the top-ranked cluster of "systems" in five language pairs: English→Arabic, English→Estonian, English→Icelandic, English→Japanese, and Japanese→English. Conversely, the reference placed relatively lowly in English→Russian (rank

9–11 of 19), English→Chinese (11–13 of 19), and Czech→German (9–12 of 21).

We extract system-level correlations for each participating auto-rater out of Table 19 and Table 20 in Appendix B in order to compute the average correlation separately per each group of language pairs. Since a strong auto-rater is more likely to outperform a weak auto-rater in *any* language pair, we compare instead the difference in correlation for the *same* auto-rater from one group to another, as a measure of its specific degradation in the face of low-quality references.

Table 9 shows the results. Indeed, our five string- and embedding-based baselines are much more sensitive to poor reference quality than any other auto-rater, by a significant margin. Likewise, on the other hand, the auto-raters that improve their performance the most on languages with poor references comprise of six of the "LLM as a Judge" models and four official submissions to the shared task: (MR7.2.1 and MR6 are based on Gemma 3 models; Q_MQM and Q_RELATIVE-MQM are based on Qwen 3.) All are reference-free systems, as expected.

4.4.5 Metric Score Difference Interpretation

Following the WMT Metrics Shared Task in the last two years, we continue to conduct analyses to find the threshold of metrics' score differences that corresponds to statistical significance of MT system rankings demonstrated by human annotators and the metrics themselves.[11] These analyses provide an interpretation of the metrics' score differences, support building an intuitive sense of metric score meanings, and encourage broader adoption of new automatic MT evaluation metrics. This year, since we expanded the number of language pairs (LPs) from 3 to 16, instead of analyzing metrics score differences by individual LP we are pooling the 14 LPs with references together in the following analyses for clear presentation and ease of understanding.

As a reminder, the results in this section should *not* be used as arguments to forego significance tests or appropriate human evaluation.

Correspondence to human scores significance: We first study the relationship between statistically significant differences in human scores and the magnitude of metric differences as in (Lo et al., 2023a). We run a one-sided paired t-test with an equal variance assumption for each system pair on segment-level human scores. After that, we fit the corresponding metric score differences and the p-values of the t-test on the human scores to an isotonic regression (Robertson et al., 1988), which predicts whether the human score difference will be significant given the metric's score difference. This year, we also consider the sign of the metric's difference. If the metric's decision disagrees with the human's but the human score difference is insignificant, we also consider that as a correct prediction. Isotonic regression produces a non-decreasing function where the classifier output can be interpreted as a confidence level.[12] We set $p_h < 0.05$ as the significance level of human scores. Thus, the output of the isotonic regression function can be viewed as $Pr(p_h < 0.05 \mid \Delta m)$ where p_h is the p-value of the t-test on the human scores for each system pair and Δm is the metric score difference.

Figure 2 shows the (log) p-value of one-sided paired t-test on the human scores against the corresponding BLEU, YISI-1, and TASER-REF score difference for each system pair. Additional figures (Figures 9-11 in Appendix C) show the same analyses for all metrics. For each metric, we can choose a particular level of confidence (i.e., a point along the y-axis on the right) to get the metric score difference cutoffs (i.e. a point along the x-axis) that this metric difference reflects significant human score differences. Drawing a horizontal line from the confidence level, say 80%, to the red line enables us to find the minimum metric difference cutoff required at the corresponding x-value down from the red line, i.e. 3.0 for BLEU in Figure 2. Using this lookup method, Table 10 show the cutoffs of Δm when $Pr(p_h < 0.05 \mid \Delta m) = 0.8$ for each metric. We run 10-fold cross-validation, and Table 10 shows that the range of precision in the cross-validation is consistently high across metrics. This means the metric cufoffs we find using the regression model are reliable.

Table 10 serves as a reference for understanding the score differences between MT systems provided by modern metrics. For example, we see that a BLEU difference of 3.0 corresponds to 80% confidence that two MT systems ranked by BLEU will match the decision made by human annotators with a significant difference. Meanwhile, a

[11] This section uses the term "metric," but the analysis is extended to auto-raters of all types as defined earlier in this paper.

[12] scikit-learn.org/stable/modules/isotonic.html

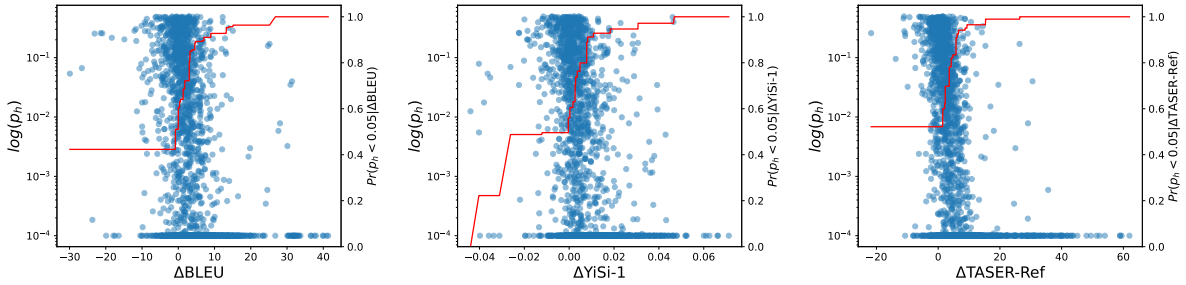

Figure 2: Log p-value of one-sided paired t-test on human scores (p_h) against each metric (left: BLEU, center: YISI-1, right: TASER-REF) score difference for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \,|\, \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001.

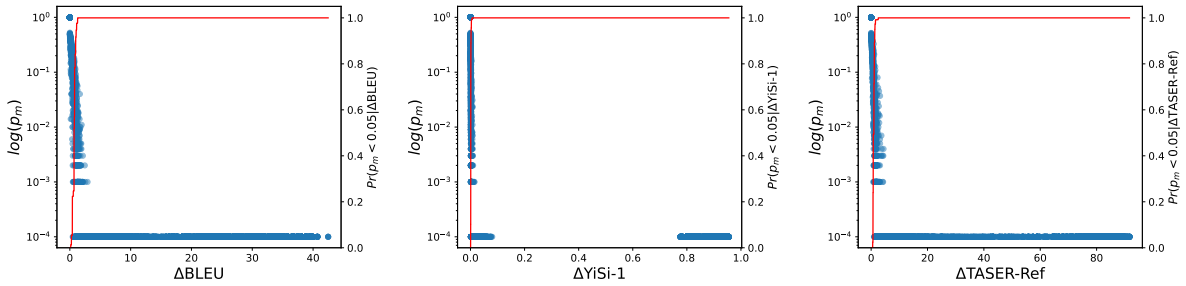

Figure 3: Log p-value of significance test with bootstrap resampling (p_m) on system-level metric scores against each metric (left: BLEU, center: YISI-1, right: TASER-REF) score difference for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \,|\, \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001.

TASER-REF score difference of 4.2 would have the same 80% chance of human-judged significant difference.

Correspondence to metric scores significance: We run a study similar to that above, but on the relations between statistically significant differences in metric scores and the magnitude of metric differences as inspired by Marie (2022). Instead of the one-sided t-test on human scores, the p-values are now obtained by running statistical significance tests with bootstrap resampling on the metric scores for each system pair. We fit the corresponding metric score differences and the p-values of the significance test to an isotonic regression for predicting whether the translation quality improvement as indicated by the metric will be significant given the metric score difference. We set $p_m < 0.05$, and thus the output of the isotonic regression function is now $Pr(p_m < 0.05 \,|\, \Delta m)$, where p_m is the p-value of the significance test on the metric scores for each system pair and Δm is the metric score difference.

Figure 3 shows the (log) p-value of the significance test with bootstrap resampling on the metric scores for BLEU, YISI-1 and TASER-REF score difference of each system pair. Additional figures (Figures 12-14 in Appendix C) show the same analyses for all metrics. Using the same lookup method described in the previous study, Table 11 shows the cut-offs of Δm when $Pr(p_m < 0.05 \,|\, \Delta m) = 0.8$ for each metric. We run 10-fold cross-validation, and Table 11 shows that the range of precision in the cross-validation is consistently high across metrics. This means that the metric cutoffs we find using the regression model are reliable.

Table 11 serves as a reference of metric differences that correspond to statistical significance with high confidence. For example, we see that a BLEU difference of 0.87 corresponds to 80% confidence that the difference is statistically significant. Meanwhile, a TASER-REF score difference of 1.1 would have the same 80% chance of being statistically significant. Our results, agreeing with Marie (2022), show that to claim significant differences ($p_m < 0.05$) on BLEU with high confidence (80%), the differences should be higher than the shared understanding (0.5 BLEU) in the research community.

We have to emphasize again that this result

Metric	minΔm	c.v. precision
Baselines		
YiSi-1	0.0078	[86-94%]
chrF	2.8	[89-93%]
spBLEU	4.0	[88-94%]
BERTScore	0.014	[87-94%]
BLEU	3.0	[86-92%]
COMET22	0.017	[84-92%]
sentinel-cand	0.23	[79-91%]
COMETKiwi22	0.048	[82-100%]
sentinel-src	—	—
Primary		
GEMBA-v2	2.1	[89-94%]
TASER-No-Ref	5.1	[93-97%]
rankedCOMET	0.057	[80-90%]
MetricX-25	2.9	[85-93%]
mr7_2_1	2.6	[86-93%]
SEGALE-QE	4.0	[83-97%]
Polycand-2	2.9	[82-93%]
Q_Relative-MQM	7.4	[87-94%]
EnsembleSlick	0.070	[55-100%]
hw-tsc	0.051	[83-100%]
UvA-MT	0.53	[74-100%]
Secondary		
TASER-Ref	4.2	[90-96%]
MetricX-25-Ref	2.3	[84-93%]
baseCOMET	0.017	[84-92%]
MetricX-25-QE	2.4	[84-91%]
mr6	2.2	[85-94%]
Q_MQM	1.9	[85-93%]
Polyic-3	3.2	[83-95%]
AutoLQA	0.015	[79-91%]
Polycand-1	2.6	[77-94%]
CollabPlus	0.025	[71-90%]
CollabSlick	0.043	[72-94%]
hw-tsc-max	0.061	[87-100%]
hw-tsc-base	0.052	[79-100%]
LLM-as-a-judge		
GPT-4_1	6.1	[89-95%]
CommandA	2.8	[86-93%]
Claude-4	3.7	[88-96%]
DeepSeek-V3	1.9	[86-95%]
Qwen3-235B	3.8	[89-95%]
Qwen2_5-7B	1.2	[82-91%]
AyaExpanse-32B	1.7	[83-95%]
Llama-3_1-8B	2.3	[75-95%]
Llama-4-Maverick	0.37	[79-91%]
CommandR7B	1.2	[76-92%]
Mistral-7B	3.5	[70-100%]
AyaExpanse-8B	0.77	[79-95%]

Table 10: Minimum Δm when $Pr(p_h < 0.05 \mid \Delta m) = 0.8$ for each metric in all language pairs with references (rounded to 2 significant figures), and the range of precision for the isotonic regression model in 10-fold cross-validation.

Metric	minΔm	c.v. precision
Baselines		
YiSi-1	0.0013	[98-100%]
chrF	0.66	[99-100%]
spBLEU	0.75	[99-100%]
BERTScore	0.0029	[99-100%]
BLEU	0.87	[99-100%]
COMET22	0.0041	[99-100%]
sentinel-cand	0.039	[99-100%]
COMETKiwi22	0.0046	[99-100%]
sentinel-src	0.00	[100-100%]
Primary		
GEMBA-v2	0.71	[99-100%]
TASER-No-Ref	1.2	[100-100%]
rankedCOMET	0.018	[100-100%]
MetricX-25	0.64	[99-100%]
mr7_2_1	0.82	[98-100%]
SEGALE-QE	0.95	[99-100%]
Polycand-2	0.47	[98-99%]
Q_Relative-MQM	2.1	[99-100%]
EnsembleSlick	0.0064	[99-100%]
hw-tsc	0.0060	[99-100%]
UvA-MT	0.030	[99-100%]
Secondary		
TASER-Ref	1.1	[99-100%]
MetricX-25-Ref	0.52	[99-100%]
baseCOMET	0.0041	[99-100%]
MetricX-25-QE	0.45	[99-100%]
mr6	0.85	[99-100%]
Q_MQM	0.53	[99-100%]
Polyic-3	0.43	[98-100%]
AutoLQA	0.0099	[99-100%]
Polycand-1	0.32	[98-100%]
CollabPlus	0.0079	[98-100%]
CollabSlick	0.0066	[99-100%]
hw-tsc-max	0.0056	[99-100%]
hw-tsc-base	0.0057	[99-100%]
LLM-as-a-judge		
GPT-4_1	1.5	[99-100%]
CommandA	0.85	[99-100%]
Claude-4	1.1	[99-100%]
DeepSeek-V3	0.61	[98-100%]
Qwen3-235B	0.87	[99-100%]
Qwen2_5-7B	0.85	[98-100%]
AyaExpanse-32B	0.55	[98-100%]
Llama-3_1-8B	1.6	[99-100%]
Llama-4-Maverick	0.74	[99-100%]
CommandR7B	0.90	[99-100%]
Mistral-7B	0.94	[98-100%]
AyaExpanse-8B	0.45	[98-100%]

Table 11: Minimum Δm when $Pr(p_m < 0.05 | \Delta m) = 0.8$ for each metric in all language pairs with references (rounded to 2 significant figures), and the range of precision for the isotonic regression model in 10-fold cross-validation.

should *not* be interpreted as evidence to forego significance test or appropriate human evaluation. Instead, we are only providing assistance to build an intuition on the meaning of the scores provided by the new metrics to encourage the transition away from lexical metrics towards more recent and stronger metrics.

5 Task 2: Span-Level Error Detection

This section presents the span-level error detection task in more detail. We discuss in more depth the error annotations per language pair (Section 5.1), and describe the baselines (Section 5.2) and the participant submissions (Section 5.3). Our meta-evaluation is described in Section 5.4. We then present the results, mostly focusing on the primary submissions, in Section 5.5.

5.1 Error Annotations

We use the ESA and MQM annotations sourced from the General MT task for this task as well, considering only the subset of documents and systems that were human-evaluated. We note that the error span patterns vary significantly per language as shown in Figure 4, which is a complementary view of Figure 1. For the translation pairs referring to lower-resource languages (e.g. English-Maasai), we frequently have the phenomenon where the whole segment is annotated as an error (frequently corresponding to hallucinated text). In contrast, annotations for higher-resource language pairs (e.g. Czech-German, English-Italian) correspond mostly to smaller, isolated error spans.

5.2 Baselines

XCOMET (Guerreiro et al., 2024) XL (3.5B) and XXL (10.7B) are neural models that are trained to identify MQM error spans in sentences along with a final quality score, thus leading to an explainable neural auto-rater. It adopts a unified input and output approach, allowing the prediction of translation quality assessment in multiple input modes (SRC-ONLY, SRC+REF and REF-ONLY), as well as generates sentence-level and word-level quality assessments. We use the SRC+REF mode with word-level predictions as the official shared task baselines.

Human2 For a subset of languages (except JA-ZH_CN and EN-KO_KR), the General MT shared task also collected a second round of human annotations. While not strictly a baseline, comparing submissions against a HUMAN2 set of evaluations provides additional insights into how automated metrics perform relative to human judgment. We report some statistics on HUMAN2 against HUMAN1 annotations in Table 12.

	# Errors	# Major Errors
Human1	33.56%	18.54%
Human2	32.48%	16.95%

Table 12: Translation error distribution on human annotations.

5.3 Submissions

We note that, this year, all task participants employed an LLM-based auto-rater to produce the fine-grained annotations. Specifically, the following systems were submitted to the task:

AIP (Yeom et al., 2025) The participants propose a tagged span annotation (TSA) approach, i.e., using reasoning LLMs to introduce inline numbered tags (e.g. $< v_0 > error\_span < \backslash v_0 >$) that explicitly mark error spans, and can easily map to diverse annotations (error severity, type, etc.) within the translated text. They enhance the tag schema to allow for annotation of omissions using zero-length tags. To be able to insert such tags on the hypothesis segments, they employed the OpenAI o3 and o4-mini reasoning models, leveraging the structured-output response mode to detect translation errors at the span level, formatted as JSON strings with the TSA approach mentioned above. They use few-shot examples and optimize for precision and minimality, explicitly prompting the models to (i) only label spans that it is confident are erroneous, and (ii) restrict the annotation to the minimal substring responsible for the error.

AutoLQA (Hrabal et al., 2025) The participants leveraged their Automatic Linguistic Quality Assessment (AutoLQA) systems, i.e., LLM-based evaluators designed to produce complete MQM-style annotations, including error spans, categories, and severities. The team fine-tuned GPT-4.1 and GPT-4o-mini model variants using internal data (≈100,000 segments), and using the WMT-QE-22 and Google-MQM datasets (dev + test) to determine performance improvements. They experiment with different prompts, controlling for the annotation structure, i.e., relaxing the MQM annotations to remove the category and approximate the ESA style. Their primary submission corresponds to the fine-tuned GPT-4o-mini and the secondary to the GPT-4.1-mini version, respectively.

GemSpanEval (Juraska et al., 2025) The participants fine-tuned a Gemma 3 27B model on past WMT MQM annotations formatted as JSON. Train-

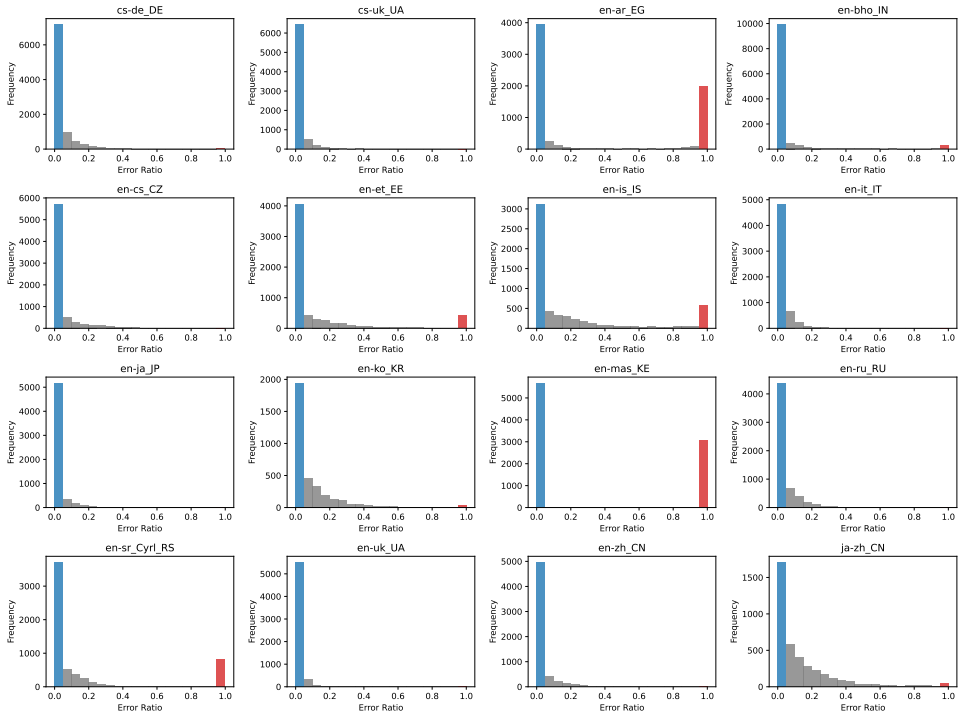

Figure 4: Distribution of error span ratio over the full segment length per language pair.

Language Pair	Baselines						Primary Submissions									Human2		
	XCOMET-XL			XCOMET-XXL			AutoLQA			AIP			GemSpanEval					
	P	R	f1	P	R	f1	P	R	f1	P	R	f1	P	R	f1	P	R	f1
CS-DE_DE	24.55	5.15	8.52	25.17	5.22	8.65	17.71	4.02	6.56	11.94	20.37	15.06	28.89	6.02	9.96	30.46	41.08	34.98
CS-UK_UA	25.02	4.00	6.90	28.21	3.56	6.32	20.73	1.93	3.54	13.60	10.16	11.63	35.94	3.58	6.52	27.67	28.95	28.30
EN-AR_EG	11.57	19.53	14.54	8.48	17.92	11.51	11.45	22.95	15.28	2.51	30.41	4.63	19.23	22.18	20.60	79.61	76.37	77.96
EN-BHO_IN	22.87	3.47	6.02	33.46	3.40	6.17	15.55	3.48	5.69	9.38	8.19	8.74	28.40	3.68	6.52	61.31	54.03	57.44
EN-CS_CZ	16.06	6.02	8.76	22.67	6.87	10.55	14.22	4.12	6.39	7.27	15.30	9.85	24.20	6.38	10.10	14.40	24.86	18.24
EN-ET_EE	14.93	16.66	15.75	16.56	17.07	16.81	14.84	11.82	13.16	5.71	20.34	8.92	23.09	12.28	16.04	33.31	32.87	33.09
EN-IS_IS	22.41	27.72	24.78	29.10	28.30	28.70	8.85	19.68	12.21	9.15	35.69	14.57	25.90	19.58	22.30	36.44	40.10	38.18
EN-IT_IT	30.60	4.16	7.33	23.40	5.40	8.77	17.89	2.55	4.47	10.45	13.70	11.86	33.71	5.47	9.41	30.52	30.62	30.57
EN-JA_JP	13.97	3.67	5.81	14.85	3.69	5.92	22.70	2.30	4.18	8.88	11.72	10.10	28.47	3.32	5.94	10.61	13.93	12.04
EN-KO_KR	8.74	14.64	10.95	9.96	17.09	12.58	20.23	7.26	10.69	4.81	25.89	8.12	17.65	10.54	13.20	-	-	-
EN-MAS_KE	10.23	34.84	15.81	11.35	36.31	17.29	15.14	38.95	21.80	27.94	28.65	28.29	35.03	35.67	35.35	94.73	92.14	93.41
EN-RU_RU	16.70	8.59	11.34	17.95	8.48	11.52	13.77	3.84	6.01	8.74	16.77	11.49	28.28	6.49	10.55	25.28	27.56	26.37
EN-SR_CYRL	21.18	21.59	21.38	24.17	21.07	22.52	13.61	15.16	14.35	6.81	27.11	10.88	21.66	15.67	18.18	61.67	58.32	59.95
EN-UK_UA	21.84	2.98	5.25	27.31	3.20	5.73	15.48	1.32	2.43	12.63	6.98	8.99	37.01	2.19	4.13	34.76	39.28	36.88
EN-ZH_CN	22.62	3.80	6.50	19.45	4.14	6.83	13.08	2.92	4.78	7.72	10.43	8.87	30.02	3.37	6.07	11.84	12.82	12.31
JA-ZH_CN	26.89	21.80	24.08	24.07	20.04	21.87	14.83	13.58	14.18	8.47	41.64	14.08	25.35	17.46	20.68	-	-	-
Average	19.39	12.41	12.11	21.01	12.61	12.61	15.63	9.74	9.11	9.75	20.21	11.63	27.68	10.87	13.47	47.04†	48.31†	47.48†

Table 13: Task 2 micro-F1 (%) by language pair for all auto-raters. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

ing data covered the period WMT20–24 (Specia et al., 2020, 2021; Zerva et al., 2024), and optimization was performed with the Adafactor (Shazeer and Stern, 2018). A central focus of their approach was the resolution of error span ambiguity, ensuring that predicted spans are uniquely identifiable within the hypothesis segment. The model was trained to extend spans with additional context whenever a substring was not unique. The context expansion covers both preceding and following context and proceeds incrementally — word by word for alphabetic languages and character by character for logographic or syllabic languages such as Chinese and Japanese — until a unique substring is obtained. The model was designed to operate in both reference-based and reference-free (QE) modes, and the team submitted both variants as their primary and secondary systems, respectively.

Language Pair	Baselines		Primary Submissions			Secondary Submissions			Human2
	XCOMET-XL	XCOMET-XXL	AutoLQA	AIP	GemSpanEval	AutoLQA-4.1	AIP	GemSpanEval-QE	
CS-DE_DE	13.07	16.22	13.91	36.45	17.08	16.63	31.22	19.14	64.46
CS-UK_UA	17.49	19.37	11.44	37.74	15.67	11.99	32.48	16.33	67.55
EN-AR_EG	10.54	10.07	11.53	18.86	12.24	10.12	14.04	12.89	79.33
EN-BHO_IN	20.00	9.88	7.14	22.44	7.01	5.46	13.40	7.71	78.86
EN-CS_CZ	10.79	10.93	17.36	31.78	12.68	13.14	25.72	14.28	60.70
EN-ET_EE	11.62	13.72	12.97	24.89	10.83	11.43	18.18	11.17	64.77
EN-IS_IS	15.87	18.50	10.01	21.83	14.95	9.59	17.22	15.24	62.51
EN-IT_IT	7.65	11.41	11.51	32.03	11.92	11.93	29.34	11.91	52.23
EN-JA_JP	10.67	16.15	11.50	44.81	8.33	10.39	37.91	12.15	64.29
EN-KO_KR	12.27	14.32	14.05	26.44	11.77	15.36	24.98	13.27	-
EN-MAS_KE	49.07	49.19	27.42	36.13	31.45	26.88	15.59	31.59	96.33
EN-RU_RU	13.19	13.80	19.09	30.50	10.77	18.06	27.29	11.20	58.49
EN-SR_CYRL	14.08	15.19	18.72	23.76	12.05	16.94	15.59	12.22	64.19
EN-UK_UA	10.17	10.79	16.97	33.69	6.55	10.07	27.20	6.37	72.02
EN-ZH_CN	6.55	8.65	32.37	38.14	7.50	32.29	33.65	8.01	59.82
JA-ZH_CN	20.17	19.78	18.90	25.83	25.74	18.48	25.03	26.76	-
Average	15.20	16.12	15.93	30.33	13.53	14.92	24.30	14.39	71.60†

Table 14: Task 2 macro-F1 (%) by language pair for all auto-rater submissions. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

5.4 Meta-Evaluation

For Task 2, we use the micro-F1 score between the predicted and the gold error spans calculated at the character level as the primary metric. The score is weighted to allow for half points for correctly identified spans with misclassified severity. Compared to the previous year, instead of computing the best matching annotation for each character (Zerva et al., 2024), we compute F1 over multiple error annotations per character, allowing for separate comparisons for each overlapping error span.

More specifically, for each hypothesis, we compute the counts for the number of "major" and "minor" errors at each character index separately for both gold annotations and predictions. This results in four statistics per hypothesis: gold major counts, gold minor counts, predicted major counts, and predicted minor counts, each of length equal to the length of the hypothesis. We then calculate a true positive (TP) score by iterating through each character position and assigning full credit based to the number of overlaps between gold and predicted counts of the same severity type (major with major, minor with minor) at each character. In the case of overlapping annotations with different severity, we assign a partial credit to the *unmatched* gold counts and predicted counts at the same character position, regardless of the original severity. This allows a predicted major error to get partial credit if it aligns with a character that was part of a gold minor error, and vice-versa. These TP scores are summed across all characters and all hypotheses. Finally, precision (P), recall (R), and F1 score are calculated based on the aggregated TP, total gold counts, and total predicted counts. The complete logic can be seen in Algorithm 1.

5.5 Main Results

Table 13 presents the complete results for all evaluated systems. Below are our primary observations and findings from these performance results.

Current auto-raters fail to localize errors. Across all language pairs where HUMAN2 scores are available, there is a very large gap between the auto-raters' performance scores and the human rater scores. HUMAN2 scores range from around 12% to over 93%, while the best auto-raters rarely exceed 35%, indicating the task is very challenging for current automated methods.

There is large variation across language pairs. No single auto-rater consistently outperforms others across all language pairs. The range of scores across different language pairs suggests varying levels of difficulty for the auto-raters. For example, most systems struggle significantly with EN-UK_UA, yielding very low F1 scores. In contrast, EN-MAS_KE allows the GemSpanEval system to achieve its peak scores. On average, GemSpanEval achieves the highest micro-F1 score (13.47%). AIP follows next with decent average performance (11.63% Primary). AutoLQA systems have the lowest average scores.

Auto-raters exhibit different precision-recall tradeoffs The AIP submissions, unlike all others, seem to be obtaining higher micro-recall, at the cost of lower micro-precision, despite including precision-focused instructions in the prompt.[13] This outcome contrasts sharply with XCOMET, AutoLQA/ESA, and GemSpanEval, which tend to be more conservative, often achieving higher precision than recall, especially on difficult languages. Human2, on the other hand, shows that high performance requires excelling in both measures, a balance that the auto-rater systems currently fail to achieve.

Auto-raters show different strengths in generalization and consistency. Table 14 reports macro-F1 score for all primary and secondary submissions. Similar to the primary evaluation (macro-F1), HUMAN2 achieves the best scores across the board. Interestingly, AIP stands out as the best submission in terms of macro-F1, averaging 30.33%. This is significantly higher than AutoLQA (15.93%) and GemSpanEval (13.53%). This largely suggests that AIP can achieve good F1 scores on average across language pairs, but its overall performance on the sheer volume of errors might be hampered by poor performance on certain heavily-weighted error segments or language pairs. For example, EN-AR_EG has many segments with full spans marked as errors due to hypotheses being in the wrong dialect (see Figure 4), and the gap in micro- and macro-F1 is large (14.23%). On the other hand, GemSpanEval's micro-F1 (13.47) is very close to its macro-F1 (13.53), which suggests a more consistent performance across language pairs in terms of the number of errors.

Overall, the results suggest that precisely locating error spans remain a challenging problem for auto-rater systems.

6 Task 3: Quality-Informed Segment-Level Error Correction

The subtask received a total of 6 submissions, from 3 participants. The results depict a clear outcome in terms of the winning system.

6.1 Data and Baselines

We reduced the number of language pairs to 6, and overall data size to a total of 6,000 instances

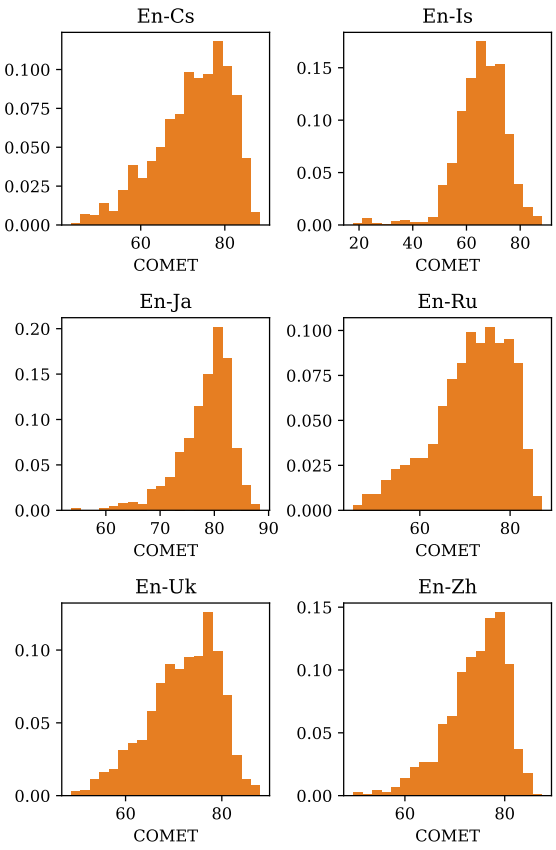

Figure 5: Task 3 - Language-pair-specific COMET score distribution of translations in the evaluation set

with an equal number of samples *per* pair. The preparation of the test set ensured representation of samples across all quartiles of the COMET score distribution. Due to Codabench limitations and memory-intensive COMET models used for evaluation, we displayed leaderboard results using a fixed 100 random samples from each submission during the competition phase. Figure 5 shows the COMET score distribution of the evaluation set of each language pair. We observe that for all language pairs, we have similar distributions skewed towards high-quality scores.

We include the following baselines along with the participants' submissions:

- **BASELINE-S1** This baseline translates the source segment from scratch using Gemma3-it-27B LLM (Gemma Team et al., 2025). It ignores the MT system output, and retranslates the input source segment. We provide the prompt used in Appendix Table 25.
- **BASELINE-S2** The final baseline uses QE information from post-edit original translation based on quality estimation from XCOMET-

[13] Table 23 shows that AIP achieves higher precision than recall on macro-F1 scores, which is aligned with their focus on being precision-focused at the instance level.

XL (Guerreiro et al., 2024), and then uses Gemma3-it-27B for APE. We provide the prompt used in Appendix Table 26.

6.2 Submissions

PHRASE (Hrabal et al., 2025) Participants leverage GPT models, specifically, o3 and o3-mini, to produce corrections over MT output given the source segment, without using the provided QE information. They use a proprietary common prompt for all systems submitted, and add variations to the prompt to change the correction strategy. The three proposed training-free approaches focus on either "only correcting errors (-S3)", "improving fluency (-S2)", and "improving fluency with steps to reason for corrections (-S1)".

SURREYPAI (Padmanabhan, 2025) Participant proposed two training-free approaches to Quality-informed error correction. The first approach (-S1) leverages the provided DA score, uses it as a selector, and routes to a specific open-weight LLM for re-translation using input QE information. This approach leverages one of six selected open-weight LLMs, wherein some LLMs were selected for their robust performance on other NLP tasks in the target language. The second approach (-S2) uses fine-grained error span information to replace an erroneous token with "__BLANK__" and then uses an LLM to replace this token contextually and "fills in the blank".

PACIFICO (Sharma, 2025) Participant proposed using natural language explanations as an intermediate step to the "detector-corrector" approach, which proposes error identification and then error correction. They use xTower to generate intermediate natural language explanations based on input QE information. The approach then feeds the explanation along with the source segment and MT output to the Gemini-1.5-Pro model to obtain final corrections.

6.3 Evaluation Metrics

We evaluate the quality of the corrections over MT, using ΔCOMET as the primary metric, and Gain-to-Edit Ratio (GER) to quantify efficiency.

ΔCOMET: Measures how much the in-post-edits improve over the original MT output (hyp) based on COMET score (Rei et al., 2022b). COMET[14] is a neural evaluation metric trained on human quality assessments, designed to capture meaning preservation and fluency by comparing translations against the source:

$$\Delta\text{COMET} = \text{COMET}(\text{src}, \text{pe}) - \text{COMET}(\text{src}, \text{hyp})$$

Positive values signal that post-editing yields a translation judged closer to human quality, while negative values imply a degradation relative to the initial MT output.

Gain-to-Edit Ratio: This metric evaluates the efficiency of edits by relating quality gains to the editing effort. According to our formulation, it is defined as the ratio between ΔCOMET and the Translation Edit Rate (TER)[15] (Snover et al., 2006) between the post-edited output (pe) and the original MT output (hyp):

$$\text{Gain-to-Edit Ratio} = \frac{\Delta\text{COMET}}{\text{TER}(\text{pe}, \text{hyp})}$$

Higher values indicate that larger quality improvements are achieved with fewer edits, while lower or negative values suggest limited or detrimental improvements relative to the editing cost.

6.4 Main Results

The main results for Task 3 are summarized in Table 15, and other metrics used for analysis are reported in Table 16. Submissions are ranked primarily by the average ΔCOMET across languages (Table 15). SURREYPAI-S1 attains the best system-wide performance, leading on both ΔCOMET for every language pair; PHRASE-S1 stands at the next best, and these two are the only submissions that surpass the BASELINE-S2 results over the primary metrics. However, in terms of efficiency of edits (GER), PHRASE-S1 obtains a higher score for En-Is, and BASELINE-S2 seems to perform the best for En-Uk.

Figure 6 illustrates the mean change in ΔCOMET for eight different systems, including two baselines, across six target languages. A clear finding is the superior performance of the SURREYPAI-S1 system, which consistently achieves a positive ΔCOMET across all language pairs, indicating an improvement in translation quality. This system shows particularly strong gains for Icelandic (`is_IS`) and Russian (`ru_RU`).

[14] Unbabel/wmt22-cometkiwi-da

[15] Computed using TERCOM-0.7.25 with default flags except the case sensitivity. Character-level tokenization is used for Chinese and Japanese, and *sacrebleu* (Post, 2018) '13a' for the rest.

System Name	En-Cs		En-Is		En-Ja		En-Ru		En-Uk		En-Zh		Average	
	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER
SURREYPAI-S1	**0.019**	**0.015**	**0.037**	0.027	**0.010**	**0.008**	**0.020**	**0.016**	**0.016**	**0.012**	**0.018**	**0.015**	**0.020**	**0.016**
PHRASE-S1	0.003	0.006	0.032	0.058	-0.006	-0.012	-0.004	-0.007	-0.003	-0.006	-0.002	-0.005	0.003	0.006
BASELINE-S2	0.000	0.000	0.007	0.026	-0.008	-0.036	0.002	0.009	0.004	0.017	-0.005	-0.023	0.000	-0.001
BASELINE-S1	-0.002	-0.002	0.008	0.005	-0.005	-0.004	-0.001	-0.001	-0.003	-0.002	0.002	0.002	0.000	0.000
PACIFICO	-0.008	-0.032	0.019	0.054	-0.018	-0.085	-0.016	-0.061	-0.008	-0.034	-0.007	-0.036	-0.006	-0.033
PHRASE-S3	-0.008	-0.030	0.027	**0.063**	-0.016	-0.060	-0.019	-0.056	-0.016	-0.045	-0.006	-0.023	-0.006	-0.025
PHRASE-S2	-0.011	-0.027	0.025	0.050	-0.018	-0.049	-0.024	-0.050	-0.020	-0.043	-0.009	-0.026	-0.010	-0.024
SURREYPAI-S2	-0.007	-0.005	-0.010	-0.006	-0.013	-0.008	-0.008	-0.005	-0.014	-0.009	-0.013	-0.010	-0.011	-0.007

Table 15: Task 3 - Performance of systems across languages with ΔCOMET and Gain to Edit Ratio (GER) metrics. Systems are ranked in order of average ΔCOMET.

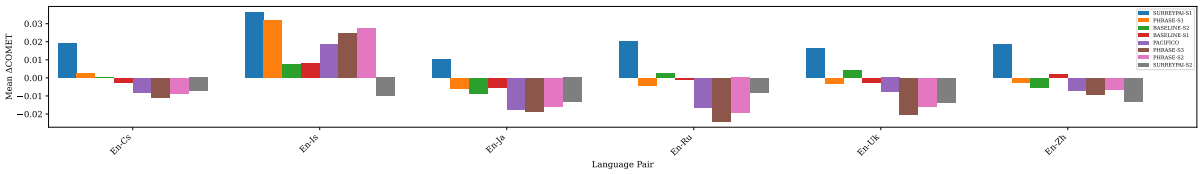

Figure 6: Task 3 - Mean ΔCOMET scores *per* **language pair** across submissions.

In contrast, most other systems exhibit mixed or negative results, with PHRASE-S2 systems frequently showing a degradation in quality, especially for Russian and Ukrainian (uk_UA). Interestingly, the English–Icelandic language pair exhibits the most notable overall improvements, with several systems achieving consistent gains. In contrast, performance on English–Japanese remains relatively limited across all systems. Correlating these outcomes with the corresponding evaluation sets suggests that LLM-based approaches follow trends observed in earlier transformer encoder–decoder APE systems (Akhbardeh et al., 2021; Bhattacharyya et al., 2023; Zerva et al., 2024). Specifically, when baseline translations are weaker—evidenced by TER distributions skewed toward higher values, APE systems tend to yield larger improvements, and vice versa. A similar pattern is observed with COMET: When the score distribution is skewed toward the upper end (Figure 5), the marginal improvements achievable by LLM-based automatic post-editing systems tend to diminish.

Figure 7 indicates that while SURREYPAI-S1 improves translations across all four domains, PHRASE-S1 shows mildly positive or negligible improvements on all. Interestingly, PACIFICO shows decent gains on *literary* and *social* domain data, but degradation in performance on the *news* and *speech* data, leads to its lower rank on the overall results. It also indicates that improvements in the *speech* domain are tough to obtain and no other system, except SURREYPAI-S1, shows improvements in terms of translation quality. We note that the speech domain data is derived from ASR transcriptions, indicating that text data derived from multimodal input may need further investigation or a different approach to correction.

Edit-Operations Figure 8 illustrates the distribution of post-editing operations like insertion (green), deletion (blue), substitution (orange), and shift (red) across various systems for six different language pairs. A clear and consistent trend is observable across all conditions: *substitution* is the most frequent edit operation, typically accounting for more than 50% of all changes. This suggests that the primary challenge for the translation systems lies in lexical choice rather than fluency. Deletion is generally the second most common error, followed by insertion. Shift operations, which correct word order, are consistently the least frequent type of edit, indicating that the models generally produce syntactically plausible translations. While this distribution pattern holds for all systems and language pairs, there are subtle variations; for instance, translations into typologically distant languages like Japanese and Chinese appear to necessitate a mildly higher proportion of insertions and deletions compared to the other language pairs.

6.5 Meta-Evaluation Metrics

While the main evaluation relies on reference-less evaluation through ΔCOMET and GER, we complement them using reference-based metrics for further analysis. In particular, we adopt ΔBLEURT (Sellam et al., 2020), a neural metric that captures semantic similarity, and

Figure 7: Task 3 - Mean ΔCOMET scores *per* **domain** across submissions.

System Name	En-Cs		En-Is		En-Ja		En-Ru		En-Uk		En-Zh		Average	
	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF/ΔchrF++
SURREYPAI-S1	-0.002	-3.460	0.053	0.000	0.003	0.000	0.009	0.000	0.002	0.000	-0.003	0.000	0.010	-0.577
PHRASE-S1	-0.012	2.256	0.031	2.856	-0.033	-0.558	-0.030	-11.636	0.023	13.758	-0.045	-0.279	-0.011	1.066
BASELINE-S2	-0.027	0.059	0.025	0.264	-0.007	-2.272	-0.006	0.733	-0.002	0.214	-0.007	-1.086	-0.004	-0.348
BASELINE-S1	-0.019	8.902	0.012	2.199	-0.009	4.358	0.011	0.954	0.018	7.267	0.007	-1.662	0.003	3.670
PACIFICO	-0.035	8.546	-0.004	10.434	-0.030	2.934	-0.039	-42.936	-0.024	6.791	-0.010	5.989	-0.024	-1.374
PHRASE-S3	-0.110	6.550	-0.075	7.117	-0.053	-0.579	-0.210	-14.537	-0.171	8.471	-0.090	0.331	-0.118	1.226
PHRASE-S2	-0.104	4.810	-0.076	5.120	-0.050	-2.228	-0.206	-16.787	-0.176	3.223	-0.083	4.760	-0.116	-0.184
SURREYPAI-S2	-0.015	-1.219	0.002	-5.248	-0.024	0.029	-0.007	0.041	0.002	-22.708	-0.024	0.000	-0.011	-4.851

Table 16: Task 3 - Performance of participant systems across languages with ΔchrF for En-Ja, En-Zh, ΔchrF++ for the rest, and ΔBLEURT for all language pairs.

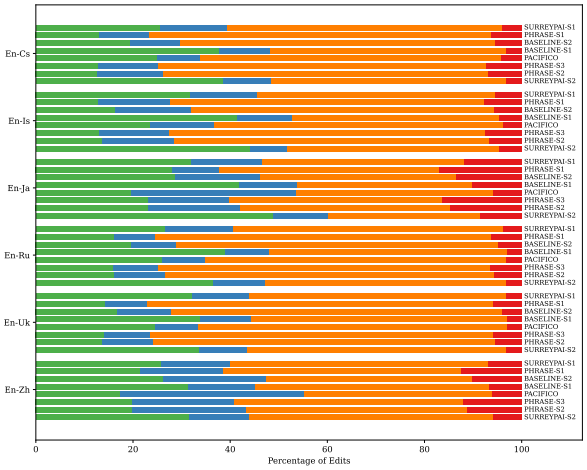

Figure 8: Task 3 - Language pair-wise **distributions of edit operations** performed on the original translation. Green indicates insertion, Blue indicates deletion, Orange indicates substitution, and Red indicates shift operations.

ΔchrF++[16] (Popović, 2017), a character n-gram overlap metric that emphasises lexical similarity. Unlike the primary reference-less metrics, which measure how much system outputs diverge from the raw translations in the desired direction, these allow us to have *an indication of how much closer system outputs have moved toward the reference or gold-standard translations* in terms of semantics and lexical distance.

Δ BLEURT: BLEURT is a learned, reference-based evaluation metric that leverages pretrained language models fine-tuned on human-rated data to capture semantic adequacy and fluency beyond surface overlap (Sellam et al., 2020). According to our formulation, ΔBLEURT measures the change in BLEURT when moving from the original MT output (hyp) to the post-edited output (pe) against the same reference (ref):

$$\Delta\text{BLEURT} = \text{BLEURT}(\text{pe, ref}) - \text{BLEURT}(\text{hyp, ref})$$

A positive ΔBLEURT indicates that resultant corrections improve semantic similarity to the reference, while a negative value suggests a reduction in quality.

ΔchrF++: chrF++ computes an F-score over word-level n-gram precision and recall between the hypothesis translation and a reference (Popović, 2017). The ΔchrF++ captures word n-gram quality gains, providing an additional cross-check beyond semantic metrics such as COMET.

$$\Delta\text{chrF++} = \quad \text{chrF++}\,(\text{pe, ref}) - \text{chrF++}\,(\text{hyp, ref})$$

A positive ΔchrF++ indicates that corrections improve similarity to the reference, while a negative value suggests a reduction in quality. We report ΔchrF for Chinese and Japanese and ΔchrF++ for the rest.

6.6 Meta-Evaluation Results

From Table 16, we observe that gains are mixed across participant systems and languages. Overall, ΔBLEURT improvements are only visible for SURREYPAI-S1 and BASELINE-S1, indicating that corrections did not substantially improve semantic adequacy across most systems. Multiple systems perform strongly on the English–Icelandic pair, particularly in ΔchrF++,

[16]Computed using sacrebleu (Post, 2018) with default flags

while English–Chinese exhibits degradation for nearly all submissions, and English–Czech results remain mixed. On English–Ukrainian, PHRASE-S1 and BASELINE-S1 clearly outperform SURREYPAI-S1 by a wide margin.

BASELINE-S1 achieves the highest system average for ΔchrF++, with PHRASE-S3 and PHRASE-S1 also showing moderate positive gains. In contrast, no other systems achieve consistent improvements in overall translation quality. Among baselines, BASELINE-S1 is comparatively closer to references, delivering competitive results across multiple languages, whereas BASELINE-S2 remains close to neutral.

Several systems (PACIFICO, PHRASE-S2, SURREYPAI-S2) show negative deltas on the secondary metrics for some languages, suggesting that their edits often diverge from reference translations despite attempted corrections. Notably, PACIFICO displays extreme variance—achieving large gains on English–Icelandic and English–Chinese, but severe degradation on English–Russian (–42.936 ΔchrF++), indicating the instability of certain LLM-based approaches across language pairs.

A closer comparison of both SurreyPAI submissions shows contrasting behaviour: SURREYPAI-S1 achieves the best system average in ΔBLEURT (+0.010), suggesting modest improvements in semantic adequacy, but its chrF average remains negative. In contrast, SURREYPAI-S2 underperforms on both metrics, with the steepest degradation in chrF (–4.851), particularly on the English–Ukrainian pair (–22.708), highlighting the sensitivity of system design choices to specific language pairs. Given the approach to SURREYPAI-S1, the system is able to retranslate and improve on translation quality, but Table 16 shows that such an approach does not bring the output closer to a known reference. At a language level, English-Icelandic seems to show the most consistent improvements, while English-Russian shows the most severe degradations. English-Chinese and English-Japanese also prove challenging, with limited gains, which may suggest that languages with morphological richness (Ukrainian and Icelandic) may offer opportunities for effective corrections, whereas typologically distant languages (Chinese and Japanese) are still harder to handle.

We also conducted batchwise significance testing with ΔCOMET scores to compare system performance. The dataset with $6,000$ instances was divided into 60 fixed batches with unique samples and 100 additional randomly sampled batches, yielding a total of 160 batchwise averages per system. For each batch, the system with the highest ΔCOMET score was identified, and the frequency of these "wins" across all batches served as an indicator of each system's consistency and robustness. Table 17 shows SURREYPAI-S1 dominates the evaluation with 157 out of 160 wins, while all other systems achieved at most two wins (PHRASE-S2 with 2 and PHRASE-S1 with 1), and the remaining systems failed to win a single batch. Testing reveals SURREYPAI-S1 to be the best system consistently across both static and random batch settings.

System	Win Count
SURREYPAI-S1	157
PHRASE-S2	2
BASELINE-S2	0
BASELINE-S1	0
PHRASE-S1	1
PACIFICO	0
PHRASE-S3	0

Table 17: Task 3 - Batch-wise meta-evaluation: winning counts per system.

6.7 Task Overview

This sub-task marks the first instance where translations were majorly generated by LLMs rather than by traditional NMT systems. We observe that LLM-based APE systems struggle to further improve these translations. This trend is analogous to earlier iterations: while neural APE systems could successfully enhance SMT outputs, they initially faced difficulties in improving NMT-generated translations. Then last year, LLM-based APE systems demonstrated the ability to improve NMT translations even for underrepresented languages. In contrast, when confronted with LLM-generated translations, even in high-resource languages, they now appear to encounter similar challenges. It requires innovative and sophisticated strategies that can effectively address the unique challenges inherent in the high-quality translations produced by LLMs.

7 Challenge Sets

For the third year, our shared task included a subtask involving challenge sets. 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.

Challenge Set	LPs	Phenomena	Items
CoDrift (Tan et al., 2025)	3	continuation drift	3,326
GAMBIT+ (Filandrianos et al., 2025)	33	gender bias	289,443
MSLC25 (Knowles et al., 2025)	2	low quality MT	369
SSA-MTE (Li et al., 2025a)	11	African languages	12,769

Table 18: Overview of the participation at the metrics challenge sets subtask.

Whereas the standard evaluation of the shared task is conducted on test sets containing generic text from real-world content, the challenge set evaluation is based on test sets designed with the aim of revealing the abilities or the weaknesses of the metrics or evaluating particular translation phenomena. In order to shed light on different perspectives on evaluation, the subtask takes place in a decentralized manner: contrary to the main metric task, the test sets are not provided by the organizers but by different research teams, who are also responsible for analyzing and presenting the results.

7.1 Subtask Structure

This subtask is made of three consecutive phases; (1) the *Breaking Round*, (2) the *Scoring Round*, and (3) the *Analysis Round*:

1. In the *Breaking Round*, every challenge set participant (*Breaker*) submits their challenge set S composed of examples for different phenomena, where every example $(s, t, r) \in S$ contains one source sentence s, one translation hypothesis t, and one reference r.

2. In the *Scoring Round*, the metrics participants from the main task (the *Builders*) are asked to score with their metrics the translations in the given test set. Also, in this phase, the metrics task organizers score all data with the baseline metrics.

3. Finally, after having gathered all metric scores, the organizers return the respective scored translations to the *Breakers* for the *Analysis Round*, where they employ their own evaluation for the performance of the metrics with regard to the phenomena they intended to test.

7.2 Challenge Set Descriptions

This year there were 4 submissions, covering a wide range of phenomena and 23 different language pairs, which supersede the official language pairs of the Metrics Shared Task. An overview of the submitted challenge sets can be seen in Table 18. A short description of every submission follows:

CoDrift (Tan et al., 2025) Quality Estimation (QE) models such as COMET-KIWI, MetricX, and ReMedy exhibit a recurring failure mode: they often assign high scores to translations that start faithfully but subsequently drift into fluent yet irrelevant content. To systematically investigate this issue, Tan et al. (2025) present CoDrift, a WMT25 challenge set designed to stress-test QE robustness against continuation drift. The dataset is constructed entirely from controlled large language model (LLM) experiments: for each source sentence, we generate multiple "drift" candidates whose continuation length and topical proximity are systematically manipulated. This design enables precise control over the degree of semantic divergence, while maintaining surface fluency, thereby creating challenging cases that can mislead current QE systems. CoDrift aims to provide the community with a targeted benchmark for diagnosing and improving QE models in the presence of subtle off-target content.

Gambit+ (Filandrianos et al., 2025) In this submission, the authors introduce GAMBIT+[17], a large-scale challenge set designed to probe gender bias in QE systems. The dataset extends the GAMBIT corpus of English gender-ambiguous occupational terms to three source languages (English, Turkish, Finnish), where occupational gender is not specified, and 11 target languages with grammatical gender: Arabic, Czech, Greek, Spanish, French, Icelandic, Italian, Portuguese, Russian, Serbian, and Ukrainian. Importantly, all occupations are linked to the ISCO-08 classification, an internationally recognized standard for categorizing jobs, which enables fine-grained per-occupation analysis and ensures coverage of the full occupational spectrum. For each source text, two parallel target translations were produced, one masculine and one feminine, differing only in the gender of the occupation and all dependent grammatical elements (e.g., pronouns, adjectives) to ensure consistency. An unbiased auto-rater should assign near-identical scores to both versions. Each source-target language pair contains over 8,500 source texts, with two parallel target translations (masculine and feminine),

[17] huggingface.co/datasets/ailsntua/gambit-plus

resulting in more than 17,000 source-translation pairs per language pair and over 550,000 pairs in total across the 33 language combinations. With its scale, full ISCO coverage, and strictly parallel design, GAMBIT+ provides a comprehensive and controlled resource for investigating gender fairness in QE metrics.

The authors benchmarked three baseline metrics and eight shared task submissions on GAMBIT+, though one baseline was excluded from the analysis since it evaluated only source texts rather than target translations. Across the remaining auto-rater systems, all showed statistically significant differences between masculine and feminine outputs, but the scale of these differences varied widely. For instance, UvA-MT and rankedCOMET displayed average normalized score gaps of over 100% and 70% respectively, while Polycand variants and Polyic metrics registered less than 4%. Bias magnitude was influenced by both the source and target languages, with English sources and target languages such as Arabic, Russian, and Icelandic exhibiting stronger disparities. At the occupational level, most auto-raters favored masculine translations overall, yet stereotypically female-associated roles (e.g., nursing, midwifery, cleaning professions) often saw the opposite pattern, reflecting known tendencies in MT systems. These results show that QE systems are sensitive to gender even in cases where they shouldn't be, amplifying occupational stereotypes rather than remaining neutral, underscoring the need for systematic auditing and fairness-aware design.

MSLC25 Challenge Set (Knowles et al., 2025) Based on the past two iterations of the Metric Score Landscape Challenge (MSLC; Lo et al., 2023b; Knowles et al., 2024), MSLC25 is a smaller-scale study of auto-rater performance on a broad range of MT quality along with several specific corner cases and phenomena. MSLC25 includes a collection of low- to medium-quality MT systems' output on Japanese–Chinese news data from the WMT25 General MT Shared Task test set. As in previous editions, the challenge set explores auto-rater scores assigned to empty strings in the source or target, showing unexpected results for some auto-rater systems. In small-scale proof-of-concept experiments (using Japanese, Chinese, English, and Czech data) the challenge set also examines auto-rater scores assigned to mixed- and wrong-language text and English language spelling variants. The results of MSLC25 continue to highlight the need for auto-rater builders to test their systems on corner cases and wide ranges of MT quality before releasing them to the broader research community.

SSA-MTE Challenge Set (Li et al., 2025b,a) The SSA-MTE challenge set is a large-scale benchmark for machine translation evaluation in Sub-Saharan African languages. It comprises 12,768 human-annotated adequacy scores across 11 language pairs involving English, French, and Portuguese, evaluated on outputs from six commercial and open-source machine translation systems. Results indicate that correlations with human judgments remain generally low, with most systems achieving Spearman correlations below the 0.4 threshold for medium-level agreement. Performance varies substantially across language pairs, and in extremely low-resource cases such as Portuguese–Emakhuwa, correlations drop to around 0.1, underscoring the challenge of evaluating MT for very low-resource African languages. Notably, the long-standing baseline metric chrF (Popović, 2015) achieves performance comparable to the strongest neural supervised submission, MetricX-25 (Juraska et al., 2025), an encoder-only regression model initialized from *Gemma3* (12B) (Gemma Team et al., 2025) and fine-tuned on WMT15–23 DA and MQM scores. However, these findings still highlight the urgent need for more robust and generalizable machine translation evaluation methods tailored to under-resourced African languages.

7.3 Challenge Set Results Overview

The studies collectively reveal critical weaknesses in current automatic MT evaluation systems. CoDrift shows that popular QE models like COMET-KIWI and MetricX often fail when translations drift into fluent but semantically irrelevant continuations, highlighting the need for robustness against subtle off-target content. GAMBIT+ uncovers systematic gender bias in QE systems across 33 language combinations, with some auto-raters showing over 100% score gaps between masculine and feminine translations, amplifying occupational stereotypes. MSLC25 emphasizes that auto-raters can behave unpredictably on low- to mid-quality outputs and corner cases such as empty strings or mixed-language outputs, stressing the importance of thorough auto-rater testing for real-world

robustness. Finally, SSA-MTE demonstrates that auto-rater correlations with human judgments remain very low for Sub-Saharan African languages, especially in extremely low-resource pairs, underscoring the urgent need for inclusive, generalizable evaluation methods.

8 Conclusion

This paper documented the results of the WMT25 shared task on automated machine translation evaluation systems, which unified the Metrics and QE Shared Tasks from previous years. The shared task this year consisted of three subtasks: (1) segment-level quality score prediction, (2) span-level translation error annotation, and (3) quality-informed segment-level error correction. Task 1 results indicate the strong performance of large LLM-as-a-judge auto-rater systems at the system level, while reference-based baseline metrics outperform LLMs at the segment level. Task 2 results indicate that accurate error detection and balancing precision and recall are persistent challenges. Task 3 results show that minimal editing is challenging even when informed by quality indicators. Robustness across the broad diversity of languages remains a major challenge across all three subtasks. As described throughout the paper, this year marked significant changes to multiple dimensions of the evaluation. Evaluation data, originating from the General-MT task, was more challenging for MT systems, and covered a diverse set of new language-pairs. The move to long segments and the adoption of ESA human annotation for most of the languages were also new. We strongly believe that these changes were all warranted by the changing landscape in the field of MT and that they better align our evaluation with the current landscape. However, these changes are also likely responsible for some of the unexpected results observed this year, particularly for Task-1. We encourage further analysis of these results by the MT research community at large.

9 Ethical Considerations

The data for this shared task was generated, screened and human-annotated by the General Machine Translation Shared Task. We acknowledge inheriting any ethical limitations and concerns raised by their shared task. We do not foresee any additional ethical concerns.

10 Acknowledgments

Results for this shared task would not be possible without the tight collaboration with the organizers of the WMT25 General MT Shared Task. We thank them for their hard work and collaboration.

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

Chrysoula Zerva was funded by the UTTER project, supported by the European Union's Horizon Europe research and innovation programme via grant agreement 101070631, by the Portuguese Recovery and Resilience Plan through projects C645008882-00000055 (Center for Responsible AI) and UID/50008: Instituto de Telecomunicações and supported by an unrestricted gift from Google (Google Research Scholar).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, pages 672–681. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2024. Together we can: Multilingual automatic post-editing for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10800–10812. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2025. Giving the old a fresh spin: Quality estimation-assisted constrained decoding for automatic post-editing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 914–925. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, Fred Blain, Tharindu Ranasinghe, and Pushpak Bhattacharyya.

2023. Quality estimation-assisted automatic post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1686–1698. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. 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.

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.

Giorgos 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*, 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, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

Sami Ul Haq and Chinonso Cynthia Osuji. 2025. Long-context Reference-based MT Quality Estimation. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. CUNI and Phrase at WMT25 MT Evaluation Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2025. GEMBA V2: Ten Judgments Are Better Than One. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate Submissions to the WMT25 Evaluation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2024. MSLC24: Further challenges for metrics on a wide landscape of translation quality. In *Proceedings of the Ninth Conference on Machine Translation*, pages 475–491. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, and Chi-kiu Lo. 2025. MSLC25: Metric Performance on Low-Quality Machine Translation, Empty Strings, and Language Variants. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025a. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025b. Preliminary ranking of WMT25 general machine translation systems. *Preprint*, arXiv:2508.14909.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494. Association for Computational Linguistics.

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

Senyu Li, Felermino Dario Mario Ali, Jiayi Wang, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, Colin Cherry, and David Ifeoluwa Adelani. 2025a. Evaluating WMT 2025 Metrics Shared Task Submissions on the SSA-MTE African Challenge Set. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Senyu Li, Jiayi Wang, Felermino DMA Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025b. 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.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023a. Beyond correlation: Making sense of the score differences of new MT evaluation metrics. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 186–199. Asia-Pacific Association for Machine Translation.

Chi-kiu Lo, Samuel Larkin, and Rebecca Knowles. 2023b. Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In *Proceedings of the Eighth Conference on Machine Translation*, pages 776–799. Association for Computational Linguistics.

Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, xiaoyu chen, and Hao Yang. 2025. HW-TSC's submissions to the WMT 2025 Segment-level quality score prediction Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Sujal Maharjan and Astha Shrestha. 2025. Ranked-COMET: Elevating a 2022 Baseline to a Top-5 Finish in the WMT 2025 QE Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Monishwaran Maheswaran, Marco Carini, Christian Federmann, and Tony Diaz. 2025. TASER: Translation assessment via systematic evaluation and reasoning. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Benjamin Marie. 2022. Yes, we need statistical significance testing. towardsai.net https://pub.towardsai.net/yes-we-need-statistical-significance-testing-927a8d21f9f0.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Govardhan Padmanabhan. 2025. Can QE-informed (Re)Translation lead to Error Correction? In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

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, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. 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.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. 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, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

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.

T. Robertson, F.T. Wright, and R. Dykstra. 1988. *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley.

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

Prashant K. Sharma. 2025. Leveraging QE-based Explanations for Quality-Informed Corrections. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

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

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91. Association for Computational Linguistics.

Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, and Toshiyuki Sekiya. 2025. CoDrift in WMT25 Metric Challenge Set Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

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.

Johnny Wei, Tom Kocmi, and Christian Federmann. 2022. Searching for a higher power in the human evaluation of MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 129–139. Association for Computational Linguistics.

Di Wu and Christof Monz. 2025. UvA-MT at WMT25 Evaluation Task: LLM Uncertainty as a Proxy for Translation Quality. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Brian Yan, Shuoyang Ding, Kuang-Da Wang, Siqi Ouyang, Oleksii Hrinchuk, Vitaly Lavrukhin, and Boris Ginsburg. 2025. Nvidia-Nemo's WMT 2025 Metrics Shared Task Submission. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and JinYeong Bak. 2025. Tagged Span Annotation for

Reasoning LLM-Based Translation Error Span Detection. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. 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.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025b. How to select datapoints for efficient human evaluation of NLG models? *Preprint*, arXiv:2501.18251.

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*, Suzhou, China. Association for Computational Linguistics.

A LLM Prompt for Task 1

Segment-level quality scores from the "LLM as a judge" submissions to Task 1 (Section 4.1.3) were prompted using the template below. Placeholders in curly braces indicate the name of the source language, name of the target language, source segment text, and target segment text.

```
Score the following translation from {source_lang} to
{target_lang} on a scale from 0 to 100, where a score of 0 means a
broken or poor translation; 33 indicates a flawed translation
with significant issues; 66 indicates a good translation with
only minor issues in grammar, fluency, or consistency; and 100
represents a perfect translation in both meaning and grammar.
Answer with only a whole number representing the score, and
nothing else.

{source_lang} source text:
{source_seg}
{target_lang} translation:
{target_seg}
```

B Complete Task 1 Results per Language Pair

Table 19 (part 1) and Table 20 (part 2) show the full detailed results of the segment-level quality score prediction task broken down by individual language pair. Correlations are computed using SPA at the system level and acc^*_{eq} at the segment level, against the "human1" gold-standard annotations, matching the approach taken in the summary results of Section 4.3.

Table 21 and Table 22 show a similar detailed per-language-pair breakdown of the results as above, except now using "human2" as the gold standard. Only language pairs annotated with ESA have this second human score; Japanese→Chinese and English→Korean are thus excluded from these tables.

Metric	cs-de		cs-uk		en-ar		en-bho		en-cs		en-et		en-is		en-it	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines																
YiSi-1	0.899 (1)	0.564 (1)	0.719 (5)	0.500 (6)	0.838 (2)	0.595 (4)	0.845 (2)	0.672 (1)	0.748 (3)	0.554 (1)	0.862 (1)	0.593 (3)	0.900 (3)	0.721 (3)	–	–
chrF	0.896 (2)	0.556 (2)	0.707 (5)	0.490 (7)	0.847 (2)	0.625 (3)	0.847 (2)	0.665 (2)	0.772 (3)	0.550 (3)	0.846 (1)	0.591 (4)	0.893 (3)	0.730 (2)	–	–
spBLEU	0.881 (3)	0.549 (3)	0.704 (5)	0.485 (8)	0.849 (2)	0.639 (2)	0.844 (2)	0.647 (3)	0.741 (4)	0.551 (3)	0.833 (2)	0.584 (6)	0.915 (2)	0.719 (3)	–	–
BERTScore	0.872 (3)	0.550 (3)	0.672 (6)	0.481 (9)	0.866 (1)	0.672 (1)	0.868 (1)	0.665 (2)	0.717 (5)	0.543 (4)	0.798 (3)	0.574 (7)	0.866 (4)	0.699 (5)	–	–
BLEU	0.873 (3)	0.540 (4)	0.661 (6)	0.475 (10)	0.853 (1)	0.628 (3)	0.872 (1)	0.644 (4)	0.734 (4)	0.539 (5)	0.794 (3)	0.567 (8)	0.879 (4)	0.693 (6)	–	–
COMET22	0.773 (6)	0.536 (5)	0.764 (4)	0.517 (4)	0.674 (5)	0.472 (7)	0.712 (5)	0.611 (6)	0.676 (6)	0.552 (3)	0.732 (4)	0.590 (5)	0.883 (4)	0.699 (5)	–	–
sentinel-cand	0.658 (8)	0.493 (10)	0.620 (6)	0.496 (6)	0.241 (13)	0.369 (13)	0.448 (12)	0.463 (15)	0.603 (8)	0.523 (7)	0.674 (6)	0.571 (7)	0.819 (6)	0.660 (9)	0.640 (7)	0.495 (5)
COMETKiwi22	0.622 (8)	0.493 (10)	0.572 (7)	0.456 (11)	0.140 (17)	0.369 (13)	0.473 (11)	0.468 (14)	0.538 (10)	0.497 (11)	0.595 (7)	0.542 (10)	0.761 (8)	0.651 (10)	0.612 (8)	0.482 (6)
sentinel-src	0.568 (9)	0.140 (23)	0.475 (8)	0.169 (26)	0.466 (8)	0.370 (12)	0.482 (11)	0.141 (28)	0.598 (8)	0.121 (27)	0.536 (9)	0.136 (25)	0.478 (10)	0.131 (30)	0.456 (10)	0.173 (22)
Primary																
GEMBA-v2	0.848 (4)	0.552 (3)	0.850 (1)	0.500 (6)	0.629 (6)	0.370 (12)	0.720 (5)	0.593 (7)	0.868 (1)	0.549 (3)	0.798 (3)	0.596 (3)	0.915 (2)	0.707 (4)	0.847 (1)	0.535 (1)
TASER-No-Ref	0.881 (2)	0.487 (11)	0.856 (1)	0.448 (12)	0.613 (7)	0.370 (12)	0.813 (3)	0.605 (6)	0.853 (1)	0.514 (8)	0.854 (1)	0.544 (10)	0.944 (1)	0.705 (4)	0.816 (2)	0.479 (6)
rankedCOMET	0.804 (5)	0.536 (4)	0.750 (4)	0.518 (3)	0.698 (4)	0.483 (6)	0.714 (5)	0.611 (5)	0.685 (5)	0.552 (2)	0.735 (4)	0.592 (4)	0.882 (4)	0.699 (5)	0.643 (7)	0.497 (5)
MetricX-25	0.773 (6)	0.539 (4)	0.823 (2)	0.528 (2)	0.364 (9)	0.370 (12)	0.647 (8)	0.552 (10)	0.744 (4)	0.551 (3)	0.751 (4)	0.597 (3)	0.858 (5)	0.692 (6)	0.728 (3)	0.539 (1)
mr7_2_1	0.853 (3)	0.471 (12)	0.847 (1)	0.450 (12)	0.247 (13)	0.370 (12)	0.659 (7)	0.523 (11)	0.828 (2)	0.450 (14)	0.764 (4)	0.485 (12)	0.832 (6)	0.607 (15)	0.826 (2)	0.431 (10)
SEGALE-QE	0.741 (6)	0.518 (6)	0.676 (5)	0.483 (8)	0.326 (10)	0.370 (12)	0.659 (7)	0.549 (10)	0.631 (7)	0.527 (6)	0.721 (5)	0.576 (7)	0.870 (4)	0.699 (5)	0.727 (3)	0.527 (2)
Polycand-2	0.694 (7)	0.503 (8)	0.687 (5)	0.498 (6)	0.250 (13)	0.369 (13)	0.521 (10)	0.480 (13)	0.621 (8)	0.528 (6)	0.733 (4)	0.572 (7)	0.878 (4)	0.680 (7)	0.682 (6)	0.502 (4)
Q_Relative-MQM	0.839 (4)	0.377 (16)	0.774 (4)	0.347 (17)	0.260 (12)	0.369 (13)	0.590 (9)	0.375 (18)	0.802 (3)	0.394 (17)	0.714 (5)	0.350 (17)	0.756 (8)	0.445 (21)	0.831 (1)	0.337 (16)
EnsembleSlick	0.718 (7)	0.484 (11)	0.646 (6)	0.457 (11)	0.332 (10)	0.369 (13)	0.538 (10)	0.468 (14)	0.629 (7)	0.496 (11)	0.728 (5)	0.541 (10)	0.833 (5)	0.625 (13)	0.715 (5)	0.493 (5)
hw-tsc	0.626 (8)	0.499 (9)	0.521 (7)	0.447 (12)	0.157 (16)	0.369 (13)	0.526 (10)	0.491 (12)	0.565 (9)	0.501 (10)	0.575 (8)	0.542 (10)	0.843 (5)	0.657 (9)	0.666 (6)	0.494 (5)
UvA-MT	0.557 (9)	0.512 (7)	0.518 (7)	0.462 (11)	0.349 (9)	0.369 (13)	0.386 (13)	0.441 (16)	0.525 (10)	0.497 (10)	0.317 (10)	0.447 (14)	0.502 (10)	0.503 (18)	0.552 (9)	0.464 (8)
Roberta-LS	–	–	–	–	–	–	–	–	0.643 (7)	0.509 (9)	–	–	–	–	–	–
Secondary																
TASER-Ref	0.917 (1)	0.553 (2)	0.874 (1)	0.512 (5)	0.710 (4)	0.432 (9)	0.797 (3)	0.651 (3)	0.872 (1)	0.557 (1)	0.867 (1)	0.610 (1)	0.940 (1)	0.736 (1)	0.851 (1)	0.524 (2)
MetricX-25-Ref	0.811 (5)	0.557 (2)	0.822 (3)	0.535 (1)	0.322 (11)	0.370 (12)	0.650 (7)	0.560 (9)	0.762 (3)	0.558 (1)	0.772 (3)	0.601 (2)	0.866 (4)	0.702 (5)	–	–
baseCOMET	0.773 (6)	0.536 (4)	0.764 (4)	0.518 (3)	0.674 (6)	0.472 (7)	0.712 (5)	0.611 (5)	0.676 (6)	0.553 (2)	0.732 (4)	0.592 (4)	0.883 (4)	0.699 (5)	0.643 (7)	0.497 (5)
MetricX-25-QE	0.727 (7)	0.520 (6)	0.752 (4)	0.511 (5)	0.338 (10)	0.370 (12)	0.661 (7)	0.565 (8)	0.692 (5)	0.541 (4)	0.732 (4)	0.589 (5)	0.833 (6)	0.678 (7)	0.710 (5)	0.539 (1)
mr6	0.809 (5)	0.470 (12)	0.846 (2)	0.438 (13)	0.285 (12)	0.370 (12)	0.655 (7)	0.518 (11)	0.795 (3)	0.437 (15)	0.759 (4)	0.463 (13)	0.824 (6)	0.599 (16)	0.798 (2)	0.424 (11)
Q_MQM	0.843 (4)	0.385 (15)	0.768 (4)	0.358 (16)	0.264 (12)	0.369 (13)	0.586 (9)	0.393 (17)	0.804 (2)	0.406 (16)	0.709 (5)	0.363 (16)	0.753 (8)	0.460 (20)	0.829 (1)	0.341 (15)
Polyic-3	0.665 (8)	0.497 (9)	0.700 (5)	0.500 (6)	0.232 (14)	0.370 (12)	0.464 (11)	0.465 (14)	0.608 (8)	0.521 (7)	0.690 (6)	0.572 (7)	0.839 (5)	0.672 (8)	0.658 (7)	0.501 (4)
AutoLQA	0.698 (7)	0.387 (15)	0.652 (6)	0.367 (15)	0.625 (6)	0.369 (13)	0.644 (8)	0.394 (17)	0.714 (5)	0.384 (18)	0.801 (3)	0.386 (15)	0.869 (4)	0.434 (22)	0.758 (3)	0.418 (11)
Polycand-1	0.679 (7)	0.499 (9)	0.645 (6)	0.490 (7)	0.241 (13)	0.378 (11)	0.484 (11)	0.465 (14)	0.613 (8)	0.524 (7)	0.700 (5)	0.565 (8)	0.839 (5)	0.670 (8)	0.666 (6)	0.497 (5)
CollabPlus	0.830 (4)	0.517 (6)	0.657 (6)	0.474 (10)	0.325 (10)	0.370 (12)	0.526 (10)	0.467 (14)	0.617 (8)	0.509 (9)	0.717 (5)	0.550 (9)	0.793 (7)	0.618 (14)	0.693 (5)	0.507 (3)
CollabSlick	0.752 (6)	0.506 (8)	0.694 (5)	0.474 (10)	0.308 (11)	0.369 (13)	0.544 (10)	0.480 (13)	0.649 (6)	0.517 (8)	0.733 (4)	0.551 (9)	0.827 (6)	0.642 (11)	0.723 (4)	0.507 (3)
hw-tsc-max	0.588 (9)	0.484 (11)	0.436 (8)	0.437 (13)	0.178 (15)	0.369 (13)	0.484 (11)	0.464 (15)	0.581 (9)	0.509 (9)	0.618 (7)	0.550 (9)	0.767 (8)	0.631 (12)	0.616 (8)	0.478 (7)
hw-tsc-base	0.588 (9)	0.484 (11)	0.436 (8)	0.437 (13)	0.145 (17)	0.369 (13)	0.484 (11)	0.464 (15)	0.541 (10)	0.491 (12)	0.531 (9)	0.521 (11)	0.767 (8)	0.631 (12)	0.616 (8)	0.478 (7)
long-context	–	–	–	–	–	–	–	–	0.639 (7)	0.510 (9)	–	–	–	–	–	–
roberta-multi	–	–	–	–	–	–	–	–	0.624 (8)	0.502 (10)	–	–	–	–	–	–
LLM-as-a-judge																
GPT-4_1	0.899 (2)	0.464 (13)	0.847 (2)	0.404 (14)	0.868 (1)	0.531 (5)	0.777 (4)	0.563 (8)	0.869 (1)	0.469 (13)	0.840 (2)	0.522 (11)	0.929 (2)	0.660 (9)	0.840 (1)	0.450 (9)
CommandA	0.859 (3)	0.406 (14)	0.833 (2)	0.354 (16)	0.737 (3)	0.439 (8)	0.694 (6)	0.341 (20)	0.871 (1)	0.377 (19)	0.783 (3)	0.350 (17)	0.808 (7)	0.429 (22)	0.831 (1)	0.397 (12)
Claude-4	0.876 (3)	0.348 (18)	0.839 (2)	0.291 (21)	0.729 (3)	0.422 (10)	0.774 (4)	0.348 (19)	0.880 (1)	0.281 (22)	0.791 (3)	0.298 (19)	0.907 (3)	0.519 (17)	0.836 (1)	0.283 (17)
DeepSeek-V3	0.857 (3)	0.364 (17)	0.854 (1)	0.331 (18)	0.554 (8)	0.369 (13)	0.639 (8)	0.379 (18)	0.848 (2)	0.307 (21)	0.846 (1)	0.330 (18)	0.862 (4)	0.473 (19)	0.826 (2)	0.338 (15)
Qwen3-235B	0.837 (4)	0.388 (15)	0.845 (2)	0.327 (19)	0.530 (8)	0.369 (13)	0.611 (9)	0.287 (22)	0.879 (1)	0.353 (20)	0.807 (3)	0.329 (18)	0.837 (5)	0.425 (23)	0.834 (1)	0.380 (13)
Qwen2_5-7B	0.721 (7)	0.362 (17)	0.657 (6)	0.320 (20)	0.339 (9)	0.369 (13)	0.535 (10)	0.313 (21)	0.779 (3)	0.307 (21)	0.541 (8)	0.292 (20)	0.659 (9)	0.378 (24)	0.817 (2)	0.368 (14)
AyaExpanse-32B	0.791 (5)	0.278 (19)	0.826 (2)	0.293 (21)	0.542 (8)	0.369 (13)	0.635 (8)	0.269 (23)	0.834 (2)	0.218 (25)	0.631 (7)	0.153 (24)	0.649 (9)	0.215 (28)	0.822 (2)	0.278 (18)
Llama-3_1-8B	0.747 (6)	0.267 (20)	0.705 (5)	0.250 (23)	0.202 (15)	0.369 (13)	0.577 (9)	0.197 (25)	0.738 (4)	0.223 (24)	0.629 (7)	0.252 (21)	0.740 (8)	0.224 (27)	0.778 (2)	0.239 (20)
Llama-4-Maverick	0.817 (4)	0.142 (23)	0.825 (2)	0.165 (27)	0.587 (7)	0.369 (13)	0.657 (7)	0.186 (26)	0.655 (6)	0.053 (28)	0.751 (4)	0.056 (27)	0.783 (7)	0.244 (25)	0.736 (3)	0.113 (23)
CommandR7B	0.769 (6)	0.218 (21)	0.708 (5)	0.214 (24)	0.171 (15)	0.369 (13)	0.549 (10)	0.181 (27)	0.682 (5)	0.186 (26)	0.465 (9)	0.195 (23)	0.442 (11)	0.186 (29)	0.803 (2)	0.258 (19)
Mistral-7B	0.639 (8)	0.282 (19)	0.560 (7)	0.276 (22)	0.207 (14)	0.369 (13)	0.476 (11)	0.221 (24)	0.618 (8)	0.233 (23)	0.487 (9)	0.222 (22)	0.463 (10)	0.238 (26)	0.630 (7)	0.258 (19)
AyaExpanse-8B	0.760 (6)	0.188 (22)	0.658 (6)	0.196 (25)	0.336 (10)	0.369 (13)	0.487 (11)	0.200 (25)	0.703 (5)	0.124 (27)	0.455 (9)	0.111 (26)	0.440 (11)	0.114 (31)	0.716 (4)	0.185 (21)

Table 19: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human1" as the gold standard. Metrics are ordered by category. Part 1 of 2.

Metric	en-ja		en-ko		en-mas		en-ru		en-sr		en-uk		en-zh		ja-zh	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines																
YiSi-1	0.836 (1)	0.534 (1)	0.839 (3)	0.470 (4)	–	–	0.557 (6)	0.503 (8)	0.864 (2)	0.604 (1)	0.645 (5)	0.515 (6)	0.643 (7)	0.486 (5)	0.877 (2)	0.503 (1)
chrF	0.813 (1)	0.525 (2)	0.852 (2)	0.465 (5)	–	–	0.542 (6)	0.492 (10)	0.840 (4)	0.589 (3)	0.677 (4)	0.520 (5)	0.658 (6)	0.479 (6)	0.857 (2)	0.487 (3)
spBLEU	0.808 (1)	0.520 (3)	0.818 (3)	0.464 (5)	–	–	0.597 (5)	0.504 (8)	0.858 (3)	0.590 (3)	0.648 (5)	0.510 (7)	0.638 (7)	0.472 (7)	0.837 (3)	0.486 (3)
BERTScore	0.818 (1)	0.521 (2)	0.805 (4)	0.466 (5)	–	–	0.583 (5)	0.498 (9)	0.838 (4)	0.596 (2)	0.604 (7)	0.505 (8)	0.621 (8)	0.482 (6)	0.853 (3)	0.491 (2)
BLEU	0.803 (1)	0.512 (4)	0.827 (3)	0.466 (5)	–	–	0.540 (6)	0.489 (10)	0.844 (4)	0.582 (4)	0.638 (6)	0.515 (6)	0.632 (7)	0.475 (7)	0.837 (3)	0.482 (4)
COMET22	0.462 (9)	0.493 (6)	0.806 (4)	0.474 (3)	–	–	0.557 (6)	0.514 (7)	0.784 (6)	0.582 (4)	0.689 (4)	0.540 (2)	0.652 (6)	0.496 (4)	0.766 (4)	0.465 (5)
sentinel-cand	0.322 (11)	0.444 (10)	0.638 (6)	0.457 (7)	0.545 (6)	0.489 (6)	0.551 (6)	0.506 (8)	0.731 (9)	0.546 (8)	0.638 (6)	0.526 (5)	0.614 (8)	0.470 (7)	0.447 (8)	0.388 (13)
COMETKiwi22	0.470 (9)	0.466 (8)	0.507 (8)	0.459 (7)	0.422 (10)	0.489 (7)	0.448 (7)	0.470 (12)	0.736 (8)	0.554 (7)	0.583 (7)	0.503 (8)	0.477 (9)	0.454 (9)	0.448 (8)	0.389 (13)
sentinel-src	0.530 (7)	0.157 (22)	0.523 (7)	0.460 (7)	0.546 (6)	0.493 (5)	0.598 (5)	0.143 (24)	0.362 (13)	0.165 (22)	0.536 (7)	0.149 (24)	0.513 (9)	0.179 (23)	0.458 (8)	0.235 (22)
Primary																
GEMBA-v2	0.712 (3)	0.512 (4)	0.901 (1)	0.471 (4)	0.682 (1)	0.489 (6)	0.775 (2)	0.551 (1)	0.875 (2)	0.587 (3)	0.773 (2)	0.532 (4)	0.832 (2)	0.487 (5)	0.858 (2)	0.460 (6)
TASER-No-Ref	0.736 (2)	0.447 (10)	0.915 (1)	0.466 (5)	0.604 (3)	0.507 (4)	0.812 (1)	0.496 (9)	0.881 (1)	0.559 (6)	0.786 (1)	0.478 (11)	0.807 (2)	0.443 (10)	0.912 (1)	0.423 (8)
rankedCOMET	0.458 (9)	0.493 (5)	0.829 (3)	0.458 (7)	0.540 (6)	0.489 (7)	0.568 (5)	0.515 (6)	0.768 (7)	0.582 (4)	0.697 (4)	0.539 (3)	0.665 (6)	0.497 (3)	0.775 (4)	0.465 (6)
MetricX-25	0.522 (8)	0.496 (5)	0.852 (2)	0.456 (9)	0.437 (10)	0.489 (6)	0.680 (3)	0.545 (2)	0.812 (5)	0.583 (4)	0.693 (4)	0.543 (2)	0.716 (5)	0.504 (2)	0.719 (5)	0.470 (5)
mr7_2_1	0.651 (5)	0.443 (10)	0.878 (1)	0.469 (4)	0.579 (4)	0.489 (7)	0.774 (2)	0.459 (13)	0.805 (5)	0.518 (11)	0.802 (1)	0.438 (12)	0.857 (1)	0.436 (11)	0.846 (3)	0.427 (8)
SEGALE-QE	0.543 (7)	0.479 (7)	0.736 (5)	0.459 (7)	0.537 (6)	0.489 (7)	0.565 (6)	0.510 (7)	0.722 (9)	0.529 (10)	0.717 (3)	0.525 (5)	0.649 (7)	0.488 (5)	0.603 (6)	0.412 (10)
Polycand-2	0.364 (11)	0.467 (8)	0.732 (5)	0.466 (5)	0.516 (7)	0.489 (7)	0.607 (4)	0.526 (4)	0.771 (7)	0.562 (6)	0.648 (5)	0.531 (4)	0.644 (7)	0.482 (6)	0.612 (6)	0.425 (8)
Q_Relative-MQM	0.732 (2)	0.386 (14)	0.811 (3)	0.460 (7)	0.542 (6)	0.489 (7)	0.734 (2)	0.368 (17)	0.860 (2)	0.450 (14)	0.764 (2)	0.354 (16)	0.831 (2)	0.381 (14)	0.848 (3)	0.414 (10)
EnsembleSlick	0.447 (10)	0.472 (8)	0.669 (6)	0.467 (5)	0.504 (7)	0.489 (6)	0.566 (6)	0.513 (7)	0.401 (12)	0.430 (15)	0.640 (6)	0.494 (9)	0.650 (7)	0.471 (7)	0.595 (7)	0.399 (12)
hw-tsc	0.527 (8)	0.476 (7)	0.622 (6)	0.456 (9)	0.502 (7)	0.489 (7)	0.446 (7)	0.487 (10)	0.759 (7)	0.554 (7)	0.633 (6)	0.513 (6)	0.476 (9)	0.468 (8)	0.526 (7)	0.403 (11)
UvA-MT	0.601 (6)	0.490 (6)	0.458 (9)	0.457 (8)	0.435 (10)	0.489 (6)	0.445 (7)	0.468 (12)	0.658 (10)	0.524 (10)	0.409 (9)	0.487 (10)	0.376 (11)	0.440 (10)	0.417 (8)	0.391 (13)
Roberta-LS	0.491 (9)	0.435 (12)	–	–	–	–	–	–	–	–	–	–	0.661 (6)	0.432 (11)	–	–
Secondary																
TASER-Ref	0.725 (3)	0.494 (5)	0.901 (1)	0.499 (1)	0.624 (2)	0.532 (3)	0.800 (1)	0.536 (3)	0.890 (1)	0.594 (2)	0.810 (1)	0.525 (5)	0.825 (2)	0.484 (5)	0.920 (1)	0.500 (1)
MetricX-25-Ref	0.559 (6)	0.509 (4)	0.851 (2)	0.458 (7)	–	–	0.686 (3)	0.556 (1)	0.824 (4)	0.588 (3)	0.735 (3)	0.550 (1)	0.756 (4)	0.513 (1)	0.768 (4)	0.490 (2)
baseCOMET	0.462 (9)	0.493 (5)	0.806 (4)	0.474 (3)	0.567 (5)	0.489 (7)	0.557 (6)	0.515 (6)	0.784 (6)	0.582 (4)	0.689 (4)	0.540 (2)	0.652 (7)	0.497 (3)	0.766 (4)	0.465 (6)
MetricX-25-QE	0.496 (9)	0.489 (6)	0.822 (3)	0.457 (8)	0.479 (8)	0.489 (7)	0.633 (4)	0.540 (3)	0.793 (6)	0.568 (5)	0.674 (4)	0.542 (2)	0.709 (5)	0.505 (2)	0.675 (6)	0.458 (7)
mr6	0.630 (5)	0.435 (11)	0.861 (2)	0.477 (2)	0.566 (5)	0.489 (7)	0.716 (3)	0.433 (15)	0.795 (5)	0.517 (11)	0.758 (2)	0.410 (13)	0.797 (3)	0.417 (12)	0.810 (3)	0.422 (8)
Q_MQM	0.732 (2)	0.399 (13)	0.819 (3)	0.463 (6)	0.651 (2)	0.489 (7)	0.733 (2)	0.372 (16)	0.858 (3)	0.459 (13)	0.758 (2)	0.360 (15)	0.833 (1)	0.387 (13)	0.841 (3)	0.417 (9)
Polyic-3	0.347 (11)	0.460 (9)	0.711 (5)	0.478 (2)	0.482 (8)	0.489 (7)	0.590 (5)	0.520 (5)	0.762 (7)	0.549 (8)	0.639 (6)	0.529 (4)	0.638 (7)	0.482 (6)	0.612 (6)	0.420 (9)
AutoLQA	0.536 (7)	0.391 (14)	0.762 (4)	0.481 (2)	0.436 (10)	0.489 (7)	0.702 (3)	0.441 (14)	0.808 (5)	0.446 (14)	0.724 (3)	0.393 (14)	0.735 (4)	0.347 (16)	0.629 (6)	0.357 (15)
Polycand-1	0.340 (11)	0.469 (8)	0.713 (5)	0.461 (6)	0.491 (8)	0.489 (7)	0.598 (5)	0.520 (5)	0.751 (8)	0.553 (7)	0.637 (6)	0.525 (5)	0.642 (7)	0.479 (6)	0.605 (6)	0.419 (9)
CollabPlus	0.419 (10)	0.470 (8)	0.697 (5)	0.464 (5)	0.422 (10)	0.489 (6)	0.555 (6)	0.510 (7)	0.415 (12)	0.428 (15)	0.668 (4)	0.512 (6)	0.651 (7)	0.478 (6)	0.697 (5)	0.419 (9)
CollabSlick	0.454 (9)	0.480 (7)	0.687 (5)	0.467 (4)	0.496 (7)	0.489 (6)	0.573 (5)	0.521 (4)	0.387 (13)	0.423 (16)	0.647 (5)	0.502 (8)	0.662 (6)	0.479 (6)	0.613 (6)	0.405 (11)
hw-tsc-max	0.549 (7)	0.474 (7)	0.572 (7)	0.456 (9)	0.466 (9)	0.489 (7)	0.441 (7)	0.475 (11)	0.711 (9)	0.542 (9)	0.608 (7)	0.503 (8)	0.443 (10)	0.467 (8)	0.521 (7)	0.396 (12)
hw-tsc-base	0.549 (7)	0.474 (7)	0.491 (8)	0.456 (9)	0.466 (9)	0.489 (7)	0.441 (7)	0.475 (11)	0.711 (9)	0.542 (9)	0.608 (7)	0.503 (8)	0.443 (10)	0.467 (8)	0.521 (7)	0.396 (12)
long-context	0.445 (10)	0.447 (10)	–	–	–	–	–	–	–	–	–	–	0.686 (5)	0.431 (11)	–	–
roberta-multi	0.477 (9)	0.440 (11)	–	–	–	–	–	–	–	–	–	–	0.646 (7)	0.437 (10)	–	–
LLM-as-a-judge																
GPT-4_1	0.761 (2)	0.454 (9)	0.877 (2)	0.503 (1)	0.610 (3)	0.543 (2)	0.816 (1)	0.458 (13)	0.888 (1)	0.506 (12)	0.781 (2)	0.434 (12)	0.798 (3)	0.415 (12)	0.940 (1)	0.464 (6)
CommandA	0.704 (4)	0.327 (16)	0.890 (1)	0.471 (4)	0.590 (4)	0.489 (7)	0.815 (1)	0.364 (17)	0.832 (4)	0.429 (15)	0.800 (1)	0.344 (17)	0.833 (2)	0.336 (17)	0.913 (1)	0.375 (14)
Claude-4	0.746 (2)	0.304 (17)	0.915 (1)	0.476 (3)	0.622 (2)	0.555 (1)	0.769 (2)	0.283 (20)	0.877 (2)	0.418 (16)	0.775 (2)	0.253 (20)	0.847 (1)	0.302 (19)	0.934 (1)	0.391 (13)
DeepSeek-V3	0.666 (5)	0.282 (18)	0.904 (1)	0.476 (3)	0.606 (3)	0.489 (7)	0.755 (2)	0.330 (19)	0.853 (3)	0.451 (14)	0.806 (1)	0.311 (19)	0.845 (1)	0.355 (15)	0.875 (2)	0.389 (13)
Qwen3-235B	0.699 (4)	0.337 (15)	0.888 (1)	0.472 (3)	0.589 (4)	0.489 (7)	0.799 (1)	0.361 (18)	0.841 (4)	0.415 (17)	0.809 (1)	0.327 (18)	0.863 (1)	0.358 (15)	0.812 (3)	0.398 (12)
Qwen2_5-7B	0.589 (6)	0.280 (18)	0.802 (4)	0.470 (4)	0.575 (5)	0.489 (7)	0.770 (2)	0.363 (17)	0.742 (8)	0.371 (18)	0.638 (6)	0.343 (17)	0.785 (3)	0.327 (18)	0.777 (4)	0.361 (15)
AyaExpanse-32B	0.626 (5)	0.225 (21)	0.896 (1)	0.456 (8)	0.582 (4)	0.489 (7)	0.744 (2)	0.227 (23)	0.690 (10)	0.274 (19)	0.766 (2)	0.219 (22)	0.821 (2)	0.237 (22)	0.794 (4)	0.331 (16)
Llama-3_1-8B	0.594 (6)	0.226 (21)	0.801 (4)	0.456 (9)	0.546 (6)	0.489 (7)	0.714 (3)	0.248 (21)	0.772 (6)	0.261 (20)	0.619 (6)	0.205 (23)	0.813 (2)	0.277 (20)	0.630 (6)	0.310 (18)
Llama-4-Maverick	0.652 (5)	0.098 (24)	0.818 (3)	0.463 (5)	0.602 (3)	0.489 (7)	0.645 (3)	0.070 (26)	0.736 (8)	0.180 (21)	0.650 (4)	0.064 (26)	0.800 (2)	0.117 (24)	0.841 (3)	0.261 (21)
CommandR7B	0.615 (5)	0.253 (19)	0.761 (4)	0.472 (3)	0.566 (5)	0.489 (7)	0.571 (5)	0.239 (22)	0.371 (13)	0.167 (22)	0.500 (8)	0.217 (22)	0.682 (5)	0.264 (21)	0.674 (6)	0.317 (17)
Mistral-7B	0.443 (10)	0.247 (20)	0.596 (6)	0.456 (9)	0.559 (5)	0.489 (7)	0.533 (6)	0.251 (21)	0.613 (11)	0.271 (19)	0.554 (7)	0.231 (21)	0.623 (7)	0.234 (22)	0.571 (7)	0.302 (19)
AyaExpanse-8B	0.558 (6)	0.131 (23)	0.729 (5)	0.456 (9)	0.502 (7)	0.489 (7)	0.633 (3)	0.126 (25)	0.422 (12)	0.182 (21)	0.509 (8)	0.123 (25)	0.670 (5)	0.178 (23)	0.703 (5)	0.288 (20)

Table 20: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human1" as the gold standard. Metrics are ordered by category. Part 2 of 2.

Metric	cs-de		cs-uk		en-ar		en-bho		en-cs		en-et		en-is	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines														
YiSi-1	0.863 (1)	0.506 (1)	0.696 (5)	0.465 (5)	0.830 (2)	0.599 (5)	0.903 (2)	0.667 (1)	0.790 (3)	0.567 (5)	0.853 (1)	0.623 (2)	0.886 (3)	0.706 (2)
chrF	0.856 (1)	0.501 (2)	0.684 (5)	0.455 (6)	0.839 (2)	0.628 (4)	0.858 (3)	0.641 (3)	0.815 (2)	0.567 (5)	0.853 (1)	0.621 (2)	0.867 (4)	0.705 (2)
spBLEU	0.842 (2)	0.497 (3)	0.676 (5)	0.457 (6)	0.852 (1)	0.647 (2)	0.902 (2)	0.635 (4)	0.789 (3)	0.556 (6)	0.864 (1)	0.622 (2)	0.895 (3)	0.698 (3)
BERTScore	0.840 (2)	0.496 (3)	0.644 (6)	0.448 (7)	0.870 (1)	0.678 (1)	0.922 (1)	0.650 (2)	0.760 (4)	0.556 (6)	0.827 (2)	0.606 (4)	0.856 (4)	0.681 (5)
BLEU	0.837 (2)	0.493 (4)	0.635 (6)	0.441 (8)	0.866 (1)	0.636 (3)	0.913 (1)	0.631 (5)	0.786 (3)	0.554 (6)	0.817 (2)	0.591 (6)	0.866 (4)	0.675 (6)
COMET22	0.754 (5)	0.495 (3)	0.774 (3)	0.478 (3)	0.628 (6)	0.469 (8)	0.772 (5)	0.604 (7)	0.684 (5)	0.576 (4)	0.691 (7)	0.607 (4)	0.869 (4)	0.690 (4)
sentinel-cand	0.650 (7)	0.460 (8)	0.622 (6)	0.463 (5)	0.195 (14)	0.362 (14)	0.508 (12)	0.465 (17)	0.598 (9)	0.539 (8)	0.643 (9)	0.575 (8)	0.803 (6)	0.652 (9)
COMETKiwi22	0.606 (8)	0.441 (10)	0.587 (6)	0.420 (11)	0.102 (17)	0.362 (14)	0.530 (11)	0.469 (17)	0.531 (10)	0.511 (11)	0.612 (9)	0.552 (11)	0.741 (7)	0.640 (11)
sentinel-src	0.544 (9)	0.196 (22)	0.501 (8)	0.208 (25)	0.464 (9)	0.363 (12)	0.515 (11)	0.144 (30)	0.563 (10)	0.122 (25)	0.517 (11)	0.131 (28)	0.474 (9)	0.129 (29)
Primary														
GEMBA-v2	0.818 (3)	0.494 (3)	0.878 (1)	0.471 (4)	0.624 (7)	0.365 (12)	0.763 (5)	0.589 (8)	0.839 (2)	0.579 (2)	0.846 (1)	0.617 (3)	0.916 (2)	0.690 (4)
TASER-No-Ref	0.845 (2)	0.458 (8)	0.867 (1)	0.424 (10)	0.605 (7)	0.362 (13)	0.856 (3)	0.588 (8)	0.841 (2)	0.524 (10)	0.872 (1)	0.569 (9)	0.945 (1)	0.680 (5)
rankedCOMET	0.783 (4)	0.495 (3)	0.747 (4)	0.478 (3)	0.652 (5)	0.482 (7)	0.772 (5)	0.605 (6)	0.694 (5)	0.577 (3)	0.699 (5)	0.606 (5)	0.866 (4)	0.690 (4)
MetricX-25	0.762 (5)	0.493 (3)	0.840 (2)	0.483 (2)	0.330 (10)	0.362 (13)	0.711 (7)	0.551 (11)	0.738 (4)	0.583 (2)	0.725 (5)	0.616 (3)	0.878 (4)	0.681 (5)
mr7_2_1	0.801 (3)	0.445 (10)	0.879 (1)	0.435 (9)	0.217 (14)	0.362 (13)	0.720 (6)	0.528 (13)	0.842 (1)	0.459 (14)	0.743 (4)	0.503 (14)	0.856 (4)	0.610 (15)
SEGALE-QE	0.755 (5)	0.474 (6)	0.695 (5)	0.454 (6)	0.286 (12)	0.362 (14)	0.716 (6)	0.538 (12)	0.640 (7)	0.545 (7)	0.686 (7)	0.581 (7)	0.884 (3)	0.679 (5)
Polycand-2	0.684 (7)	0.465 (7)	0.701 (4)	0.457 (6)	0.203 (14)	0.362 (14)	0.568 (10)	0.480 (16)	0.626 (8)	0.541 (7)	0.700 (5)	0.574 (8)	0.856 (4)	0.662 (8)
Q_Relative-MQM	0.803 (3)	0.386 (13)	0.795 (3)	0.368 (15)	0.254 (13)	0.362 (14)	0.644 (9)	0.393 (19)	0.812 (2)	0.414 (16)	0.728 (4)	0.365 (19)	0.777 (7)	0.437 (21)
EnsembleSlick	0.688 (7)	0.455 (9)	0.664 (5)	0.425 (10)	0.290 (11)	0.362 (14)	0.603 (9)	0.468 (17)	0.629 (8)	0.513 (11)	0.692 (6)	0.541 (12)	0.824 (5)	0.613 (14)
hw-tsc	0.610 (8)	0.452 (9)	0.540 (7)	0.411 (12)	0.115 (17)	0.362 (14)	0.586 (10)	0.498 (15)	0.570 (10)	0.522 (10)	0.583 (10)	0.554 (11)	0.840 (5)	0.646 (10)
UvA-MT	0.551 (9)	0.458 (8)	0.541 (7)	0.428 (10)	0.372 (10)	0.362 (14)	0.445 (13)	0.465 (17)	0.505 (11)	0.508 (11)	0.332 (13)	0.451 (16)	0.531 (9)	0.513 (17)
Roberta-LS	–	–	–	–	–	–	–	–	0.654 (7)	0.524 (10)	–	–	–	–
Secondary														
TASER-Ref	0.886 (1)	0.504 (2)	0.877 (1)	0.487 (2)	0.698 (4)	0.437 (10)	0.852 (3)	0.651 (2)	0.859 (1)	0.582 (2)	0.849 (1)	0.642 (1)	0.942 (1)	0.722 (1)
MetricX-25-Ref	0.793 (4)	0.510 (1)	0.844 (2)	0.494 (1)	0.289 (12)	0.362 (13)	0.699 (7)	0.550 (11)	0.754 (4)	0.593 (1)	0.742 (4)	0.621 (2)	0.873 (4)	0.686 (4)
baseCOMET	0.754 (5)	0.495 (3)	0.774 (3)	0.478 (3)	0.628 (7)	0.469 (8)	0.772 (5)	0.605 (6)	0.684 (6)	0.577 (3)	0.691 (6)	0.607 (4)	0.869 (4)	0.690 (4)
MetricX-25-QE	0.717 (6)	0.482 (5)	0.764 (3)	0.473 (3)	0.303 (11)	0.362 (13)	0.726 (6)	0.556 (10)	0.678 (6)	0.570 (5)	0.696 (6)	0.602 (5)	0.845 (5)	0.667 (7)
mr6	0.779 (4)	0.452 (9)	0.872 (1)	0.434 (9)	0.256 (13)	0.362 (14)	0.723 (6)	0.521 (14)	0.790 (3)	0.458 (14)	0.741 (4)	0.478 (15)	0.839 (5)	0.592 (16)
Q_MQM	0.805 (3)	0.394 (12)	0.787 (3)	0.379 (14)	0.258 (13)	0.362 (14)	0.642 (9)	0.407 (18)	0.812 (2)	0.422 (15)	0.723 (5)	0.377 (18)	0.777 (7)	0.451 (20)
Polyic-3	0.666 (7)	0.458 (8)	0.706 (4)	0.458 (6)	0.184 (15)	0.362 (14)	0.526 (11)	0.476 (16)	0.604 (9)	0.533 (9)	0.664 (8)	0.574 (8)	0.820 (6)	0.653 (9)
AutoLQA	0.711 (6)	0.381 (13)	0.664 (5)	0.378 (14)	0.589 (7)	0.362 (13)	0.684 (7)	0.397 (19)	0.718 (4)	0.394 (17)	0.755 (3)	0.386 (17)	0.884 (3)	0.433 (21)
Polycand-1	0.670 (7)	0.463 (7)	0.665 (5)	0.453 (7)	0.198 (14)	0.365 (12)	0.526 (11)	0.468 (17)	0.611 (8)	0.540 (7)	0.664 (8)	0.569 (9)	0.821 (6)	0.651 (9)
CollabPlus	0.790 (4)	0.480 (5)	0.669 (5)	0.438 (8)	0.282 (12)	0.362 (13)	0.592 (10)	0.466 (17)	0.624 (8)	0.537 (8)	0.682 (7)	0.558 (10)	0.777 (7)	0.615 (14)
CollabSlick	0.717 (6)	0.473 (6)	0.712 (4)	0.439 (8)	0.267 (12)	0.362 (14)	0.607 (9)	0.480 (16)	0.650 (7)	0.539 (8)	0.696 (5)	0.553 (11)	0.822 (5)	0.628 (12)
hw-tsc-max	0.580 (9)	0.438 (11)	0.450 (8)	0.405 (13)	0.138 (16)	0.362 (14)	0.546 (11)	0.479 (16)	0.592 (9)	0.532 (9)	0.614 (9)	0.563 (10)	0.753 (7)	0.621 (13)
hw-tsc-base	0.580 (9)	0.438 (11)	0.450 (8)	0.405 (13)	0.109 (17)	0.362 (14)	0.546 (11)	0.479 (16)	0.529 (11)	0.503 (12)	0.564 (11)	0.529 (13)	0.753 (7)	0.621 (13)
long-context	–	–	–	–	–	–	–	–	0.647 (7)	0.533 (9)	–	–	–	–
roberta-multi	–	–	–	–	–	–	–	–	0.633 (8)	0.526 (10)	–	–	–	–
LLM-as-a-judge														
GPT-4_1	0.845 (2)	0.435 (11)	0.852 (2)	0.405 (13)	0.847 (2)	0.547 (6)	0.823 (4)	0.570 (9)	0.868 (1)	0.491 (13)	0.851 (1)	0.549 (11)	0.944 (1)	0.645 (10)
CommandA	0.845 (2)	0.397 (12)	0.854 (2)	0.380 (14)	0.723 (3)	0.449 (9)	0.765 (5)	0.351 (22)	0.868 (1)	0.390 (17)	0.755 (4)	0.367 (19)	0.826 (5)	0.425 (22)
Claude-4	0.856 (2)	0.351 (17)	0.852 (2)	0.338 (17)	0.702 (3)	0.417 (11)	0.829 (4)	0.364 (21)	0.871 (1)	0.289 (20)	0.775 (3)	0.315 (22)	0.918 (2)	0.506 (18)
DeepSeek-V3	0.824 (3)	0.375 (15)	0.873 (1)	0.362 (16)	0.531 (8)	0.362 (14)	0.711 (7)	0.378 (20)	0.870 (1)	0.325 (19)	0.783 (3)	0.347 (20)	0.857 (4)	0.465 (19)
Qwen3-235B	0.799 (4)	0.379 (14)	0.862 (1)	0.343 (17)	0.502 (9)	0.362 (14)	0.683 (8)	0.302 (24)	0.870 (1)	0.367 (18)	0.767 (3)	0.336 (21)	0.861 (4)	0.421 (22)
Qwen2_5-7B	0.704 (6)	0.360 (16)	0.667 (5)	0.336 (18)	0.313 (11)	0.362 (13)	0.599 (10)	0.322 (23)	0.782 (3)	0.322 (19)	0.530 (11)	0.290 (23)	0.670 (8)	0.365 (23)
AyaExpanse-32B	0.777 (4)	0.300 (18)	0.863 (1)	0.323 (19)	0.517 (8)	0.362 (13)	0.689 (7)	0.276 (25)	0.850 (1)	0.225 (23)	0.588 (10)	0.157 (27)	0.687 (8)	0.208 (27)
Llama-3_1-8B	0.718 (5)	0.293 (19)	0.727 (4)	0.293 (21)	0.167 (15)	0.362 (14)	0.639 (9)	0.203 (28)	0.733 (4)	0.235 (21)	0.616 (9)	0.254 (24)	0.746 (7)	0.218 (26)
Llama-4-Maverick	0.794 (4)	0.195 (22)	0.844 (2)	0.224 (24)	0.569 (8)	0.362 (14)	0.730 (6)	0.205 (28)	0.681 (6)	0.042 (27)	0.715 (5)	0.054 (30)	0.814 (6)	0.236 (25)
CommandR7B	0.729 (5)	0.254 (20)	0.747 (3)	0.254 (22)	0.179 (15)	0.362 (14)	0.614 (9)	0.187 (29)	0.671 (6)	0.185 (24)	0.457 (12)	0.202 (26)	0.451 (10)	0.188 (28)
Mistral-7B	0.620 (8)	0.296 (18)	0.598 (6)	0.301 (20)	0.170 (15)	0.362 (14)	0.534 (11)	0.229 (26)	0.600 (9)	0.230 (22)	0.448 (12)	0.222 (25)	0.487 (9)	0.242 (24)
AyaExpanse-8B	0.740 (5)	0.231 (21)	0.673 (5)	0.239 (23)	0.310 (11)	0.362 (14)	0.555 (10)	0.217 (27)	0.742 (4)	0.117 (26)	0.454 (12)	0.107 (29)	0.465 (10)	0.116 (30)

Table 21: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human2" as the gold standard. Metrics are ordered by category. Part 1 of 2.

Metric	en-it		en-ja		en-mas		en-ru		en-sr		en-uk		en-zh	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines														
YiSi-1	–	–	0.734 (2)	0.516 (2)	–	–	0.585 (9)	0.518 (5)	0.840 (3)	0.651 (2)	0.702 (4)	0.521 (6)	0.658 (3)	0.503 (3)
chrF	–	–	0.734 (2)	0.513 (2)	–	–	0.583 (9)	0.515 (6)	0.817 (4)	0.638 (4)	0.736 (3)	0.526 (5)	0.677 (3)	0.506 (2)
spBLEU	–	–	0.711 (3)	0.508 (3)	–	–	0.632 (7)	0.521 (5)	0.847 (3)	0.642 (3)	0.705 (4)	0.517 (7)	0.657 (3)	0.503 (4)
BERTScore	–	–	0.729 (2)	0.508 (3)	–	–	0.596 (9)	0.514 (6)	0.821 (4)	0.642 (3)	0.655 (6)	0.507 (8)	0.633 (4)	0.496 (5)
BLEU	–	–	0.728 (2)	0.507 (3)	–	–	0.566 (10)	0.511 (6)	0.830 (4)	0.637 (4)	0.688 (5)	0.514 (7)	0.662 (3)	0.505 (2)
COMET22	–	–	0.576 (5)	0.498 (5)	–	–	0.571 (10)	0.516 (5)	0.835 (3)	0.654 (2)	0.698 (4)	0.539 (2)	0.584 (5)	0.503 (4)
sentinel-cand	0.665 (6)	0.507 (5)	0.479 (7)	0.468 (8)	0.443 (13)	0.673 (1)	0.570 (10)	0.496 (8)	0.781 (6)	0.603 (7)	0.633 (6)	0.511 (7)	0.483 (8)	0.477 (7)
COMETKiwi22	0.601 (7)	0.484 (8)	0.454 (8)	0.461 (9)	0.549 (10)	0.673 (1)	0.449 (11)	0.487 (10)	0.714 (8)	0.592 (8)	0.534 (8)	0.509 (8)	0.526 (6)	0.476 (7)
sentinel-src	0.422 (8)	0.172 (23)	0.524 (6)	0.157 (22)	0.547 (11)	0.673 (1)	0.556 (10)	0.142 (24)	0.455 (11)	0.155 (23)	0.563 (7)	0.155 (21)	0.551 (6)	0.161 (25)
Primary														
GEMBA-v2	0.830 (2)	0.540 (2)	0.773 (1)	0.526 (1)	0.558 (10)	0.673 (1)	0.835 (3)	0.559 (1)	0.891 (1)	0.644 (3)	0.836 (1)	0.541 (2)	0.744 (1)	0.494 (5)
TASER-No-Ref	0.807 (2)	0.486 (8)	0.781 (1)	0.454 (9)	0.696 (2)	0.673 (1)	0.852 (2)	0.503 (7)	0.890 (1)	0.618 (5)	0.841 (1)	0.477 (10)	0.721 (2)	0.451 (9)
rankedCOMET	0.663 (6)	0.502 (6)	0.573 (6)	0.499 (4)	0.615 (7)	0.673 (1)	0.583 (9)	0.516 (5)	0.829 (4)	0.655 (1)	0.704 (4)	0.539 (3)	0.594 (4)	0.504 (3)
MetricX-25	0.744 (3)	0.543 (2)	0.637 (4)	0.512 (3)	0.547 (11)	0.673 (1)	0.707 (6)	0.541 (3)	0.813 (4)	0.655 (1)	0.729 (4)	0.542 (2)	0.608 (4)	0.509 (2)
mr7_2_1	0.855 (1)	0.433 (11)	0.740 (2)	0.444 (11)	0.663 (4)	0.673 (1)	0.809 (3)	0.466 (12)	0.808 (5)	0.574 (9)	0.855 (1)	0.448 (11)	0.726 (2)	0.426 (11)
SEGALE-QE	0.743 (3)	0.536 (3)	0.531 (6)	0.476 (7)	0.636 (6)	0.673 (1)	0.596 (9)	0.514 (6)	0.666 (9)	0.561 (11)	0.714 (4)	0.528 (5)	0.589 (4)	0.486 (6)
Polycand-2	0.703 (5)	0.514 (4)	0.516 (7)	0.482 (6)	0.633 (6)	0.673 (1)	0.613 (9)	0.519 (5)	0.787 (6)	0.620 (5)	0.634 (6)	0.533 (4)	0.547 (6)	0.483 (6)
Q_Relative-MQM	0.825 (2)	0.347 (17)	0.731 (2)	0.384 (13)	0.626 (6)	0.673 (1)	0.783 (4)	0.366 (17)	0.852 (3)	0.498 (14)	0.805 (2)	0.352 (15)	0.759 (1)	0.379 (15)
EnsembleSlick	0.716 (5)	0.497 (6)	0.582 (5)	0.481 (7)	0.575 (9)	0.673 (1)	0.614 (9)	0.499 (8)	0.395 (11)	0.405 (18)	0.632 (6)	0.497 (9)	0.525 (7)	0.464 (8)
hw-tsc	0.656 (6)	0.494 (7)	0.473 (8)	0.465 (8)	0.588 (8)	0.673 (1)	0.461 (11)	0.499 (8)	0.744 (7)	0.600 (7)	0.585 (7)	0.513 (7)	0.521 (7)	0.480 (7)
UvA-MT	0.582 (7)	0.473 (9)	0.561 (6)	0.489 (6)	0.309 (14)	0.673 (1)	0.437 (11)	0.478 (11)	0.636 (9)	0.589 (8)	0.470 (9)	0.500 (9)	0.404 (9)	0.460 (8)
Roberta-LS	–	–	0.566 (6)	0.450 (10)	–	–	–	–	–	–	–	–	0.574 (5)	0.444 (10)
Secondary														
TASER-Ref	0.857 (1)	0.538 (2)	0.757 (1)	0.499 (4)	0.706 (2)	0.673 (1)	0.863 (2)	0.542 (3)	0.882 (1)	0.658 (1)	0.858 (1)	0.538 (3)	0.740 (1)	0.479 (7)
MetricX-25-Ref	–	–	0.648 (4)	0.517 (2)	–	–	0.726 (5)	0.549 (2)	0.812 (5)	0.656 (1)	0.756 (3)	0.558 (1)	0.641 (3)	0.515 (1)
baseCOMET	0.663 (6)	0.502 (6)	0.576 (6)	0.499 (4)	0.628 (6)	0.673 (1)	0.571 (10)	0.516 (5)	0.835 (4)	0.655 (1)	0.698 (4)	0.539 (2)	0.584 (5)	0.504 (3)
MetricX-25-QE	0.722 (4)	0.546 (1)	0.575 (6)	0.511 (3)	0.574 (9)	0.673 (1)	0.669 (7)	0.532 (4)	0.792 (6)	0.644 (3)	0.687 (5)	0.539 (2)	0.588 (5)	0.509 (2)
mr6	0.818 (2)	0.429 (11)	0.720 (2)	0.447 (10)	0.655 (5)	0.673 (1)	0.735 (4)	0.439 (14)	0.792 (5)	0.569 (10)	0.796 (2)	0.421 (12)	0.708 (2)	0.416 (12)
Q_MQM	0.826 (2)	0.352 (16)	0.729 (2)	0.394 (12)	0.505 (12)	0.673 (1)	0.785 (4)	0.370 (16)	0.858 (2)	0.505 (13)	0.798 (2)	0.358 (14)	0.766 (1)	0.390 (14)
Polyic-3	0.675 (6)	0.506 (5)	0.476 (8)	0.479 (7)	0.583 (8)	0.673 (1)	0.602 (9)	0.512 (6)	0.794 (5)	0.604 (7)	0.631 (6)	0.525 (6)	0.513 (7)	0.481 (6)
AutoLQA	0.774 (3)	0.426 (12)	0.691 (3)	0.398 (12)	0.577 (9)	0.673 (1)	0.733 (4)	0.426 (15)	0.815 (4)	0.466 (16)	0.760 (3)	0.392 (13)	0.586 (5)	0.333 (16)
Polycand-1	0.683 (6)	0.506 (5)	0.500 (7)	0.483 (6)	0.604 (7)	0.673 (1)	0.623 (8)	0.514 (6)	0.792 (6)	0.615 (6)	0.632 (6)	0.529 (5)	0.532 (6)	0.480 (7)
CollabPlus	0.702 (5)	0.514 (4)	0.559 (6)	0.487 (6)	0.531 (11)	0.673 (1)	0.602 (9)	0.505 (7)	0.412 (11)	0.396 (19)	0.654 (6)	0.509 (8)	0.518 (7)	0.479 (7)
CollabSlick	0.726 (4)	0.513 (4)	0.585 (5)	0.488 (6)	0.566 (10)	0.673 (1)	0.622 (8)	0.508 (7)	0.392 (11)	0.400 (19)	0.638 (6)	0.504 (8)	0.538 (6)	0.475 (7)
hw-tsc-max	0.594 (7)	0.483 (8)	0.466 (8)	0.458 (9)	0.580 (9)	0.673 (1)	0.431 (12)	0.492 (9)	0.683 (9)	0.571 (9)	0.552 (8)	0.502 (9)	0.512 (7)	0.478 (7)
hw-tsc-base	0.594 (7)	0.483 (8)	0.466 (8)	0.458 (9)	0.580 (9)	0.673 (1)	0.431 (12)	0.492 (9)	0.683 (9)	0.571 (9)	0.552 (8)	0.502 (9)	0.512 (7)	0.478 (7)
long-context	–	–	0.566 (6)	0.485 (6)	–	–	–	–	–	–	–	–	0.550 (6)	0.438 (10)
roberta-multi	–	–	0.523 (7)	0.449 (10)	–	–	–	–	–	–	–	–	0.550 (6)	0.453 (9)
LLM-as-a-judge														
GPT-4_1	0.826 (2)	0.456 (10)	0.758 (1)	0.450 (10)	0.695 (3)	0.673 (1)	0.863 (2)	0.457 (13)	0.871 (2)	0.550 (12)	0.842 (1)	0.446 (11)	0.733 (2)	0.407 (13)
CommandA	0.829 (2)	0.404 (13)	0.773 (1)	0.341 (14)	0.694 (3)	0.673 (1)	0.891 (1)	0.366 (17)	0.823 (4)	0.477 (15)	0.843 (1)	0.360 (14)	0.717 (2)	0.322 (18)
Claude-4	0.842 (1)	0.300 (18)	0.745 (2)	0.301 (16)	0.702 (2)	0.673 (1)	0.830 (3)	0.285 (20)	0.864 (2)	0.457 (17)	0.830 (2)	0.264 (18)	0.737 (1)	0.268 (20)
DeepSeek-V3	0.843 (1)	0.349 (16)	0.755 (1)	0.295 (16)	0.719 (1)	0.673 (1)	0.838 (3)	0.334 (19)	0.832 (4)	0.496 (14)	0.855 (1)	0.318 (17)	0.717 (2)	0.330 (17)
Qwen3-235B	0.843 (1)	0.386 (14)	0.730 (2)	0.333 (15)	0.666 (4)	0.673 (1)	0.863 (2)	0.357 (18)	0.825 (4)	0.459 (17)	0.856 (1)	0.337 (16)	0.744 (1)	0.338 (16)
Qwen2_5-7B	0.816 (2)	0.378 (15)	0.725 (2)	0.294 (17)	0.662 (4)	0.673 (1)	0.776 (4)	0.361 (17)	0.749 (7)	0.396 (19)	0.679 (5)	0.348 (15)	0.716 (2)	0.310 (19)
AyaExpanse-32B	0.824 (2)	0.285 (19)	0.697 (3)	0.230 (21)	0.661 (4)	0.673 (1)	0.774 (4)	0.226 (23)	0.702 (8)	0.284 (20)	0.776 (3)	0.225 (19)	0.672 (3)	0.215 (23)
Llama-3_1-8B	0.766 (3)	0.242 (21)	0.666 (3)	0.236 (20)	0.613 (7)	0.673 (1)	0.681 (6)	0.243 (21)	0.759 (7)	0.268 (21)	0.672 (5)	0.217 (20)	0.677 (3)	0.258 (21)
Llama-4-Maverick	0.759 (3)	0.116 (24)	0.700 (3)	0.101 (24)	0.660 (4)	0.673 (1)	0.689 (6)	0.071 (26)	0.733 (7)	0.175 (22)	0.619 (6)	0.070 (23)	0.691 (3)	0.089 (26)
CommandR7B	0.765 (3)	0.256 (20)	0.708 (3)	0.266 (18)	0.677 (4)	0.673 (1)	0.550 (10)	0.238 (22)	0.355 (12)	0.150 (24)	0.487 (8)	0.228 (19)	0.593 (4)	0.264 (20)
Mistral-7B	0.615 (7)	0.252 (20)	0.514 (7)	0.254 (19)	0.646 (5)	0.673 (1)	0.535 (10)	0.245 (21)	0.599 (10)	0.287 (20)	0.545 (8)	0.229 (19)	0.534 (6)	0.226 (22)
AyaExpanse-8B	0.693 (5)	0.180 (22)	0.715 (2)	0.140 (23)	0.656 (4)	0.673 (1)	0.592 (9)	0.130 (25)	0.404 (11)	0.157 (23)	0.554 (7)	0.130 (22)	0.660 (3)	0.167 (24)

Table 22: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human2" as the gold standard. Metrics are ordered by category. Part 2 of 2.

C Task 1 Score Difference Interpretation Additional Figures

Figures 9-11 show the (log) p-value of one-sided paired t-test on the human scores against the score difference of each auto-rater for each system pair. Figures 12-13 show the (log) p-value of significance test with bootstrap resampling on the auto-rater scores against the score difference of that auto-rater for each system pair in each language pair.

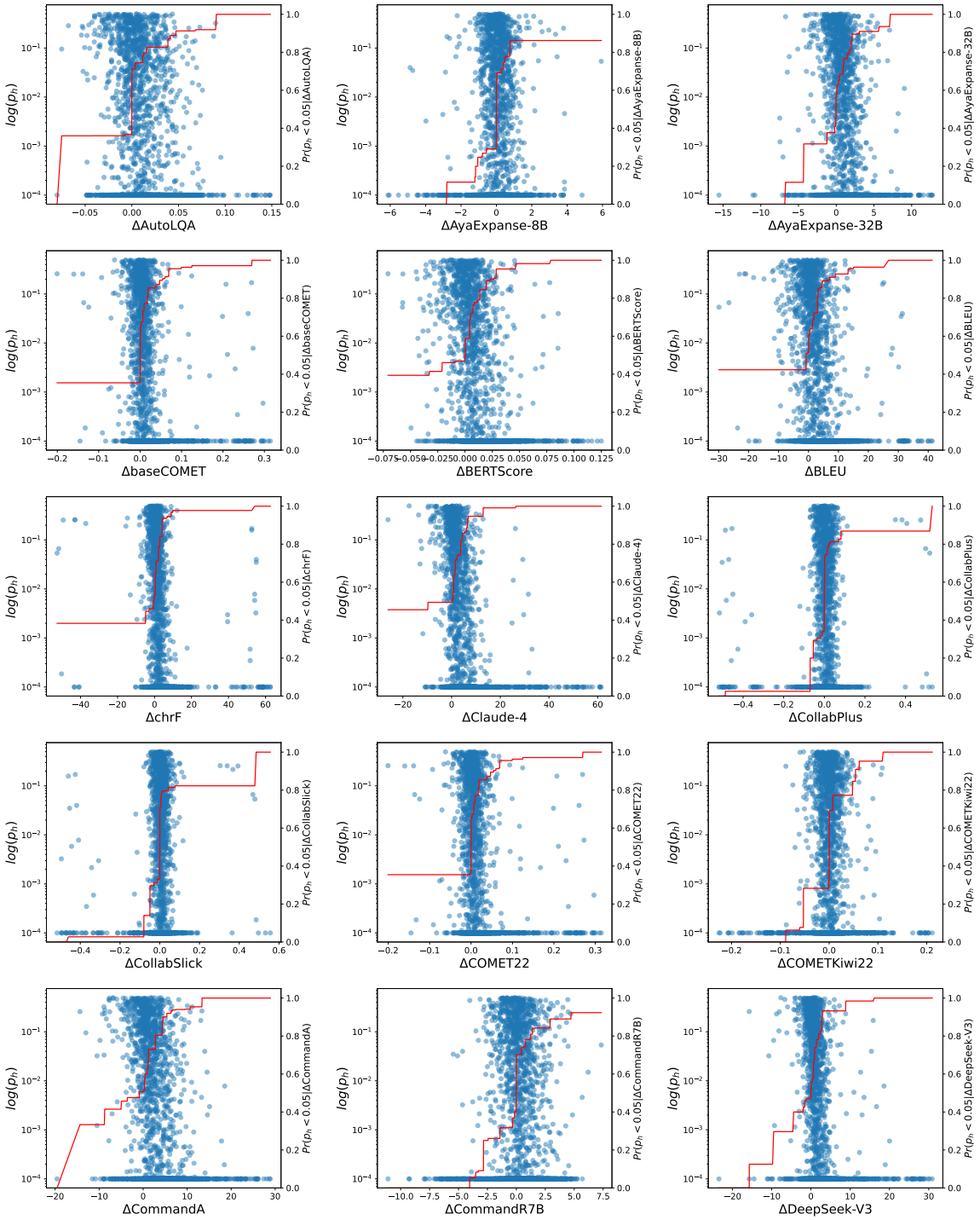

Figure 9: Log p-value of one-sided paired t-test on MQM scores (p_h) against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \mid \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001. (Part 1/3)

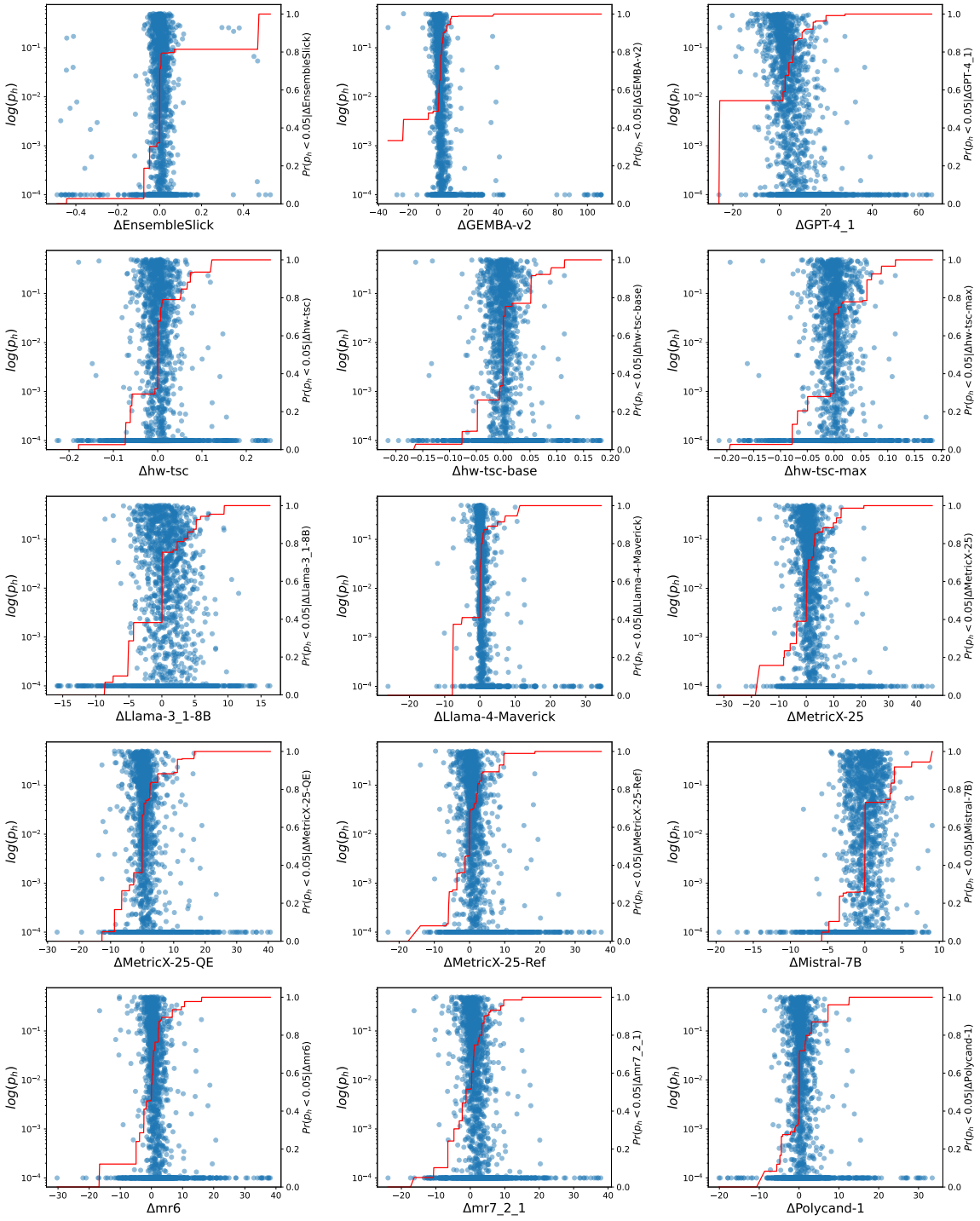

Figure 10: Log p-value of one-sided paired t-test on MQM scores (p_h) against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \mid \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001.(Part 2/3)

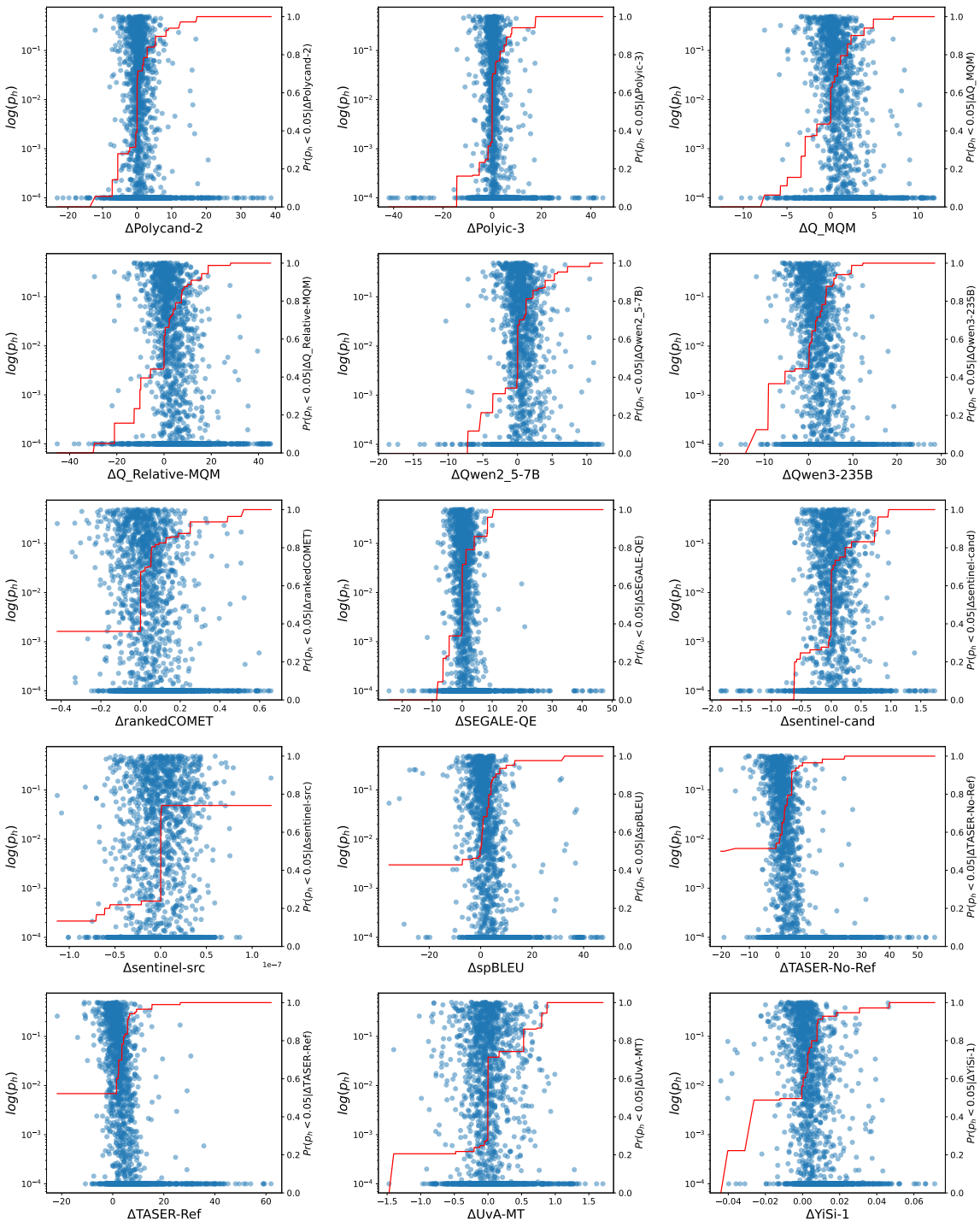

Figure 11: Log p-value of one-sided paired t-test on MQM scores (p_h) against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \mid \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001.(Part 3/3)

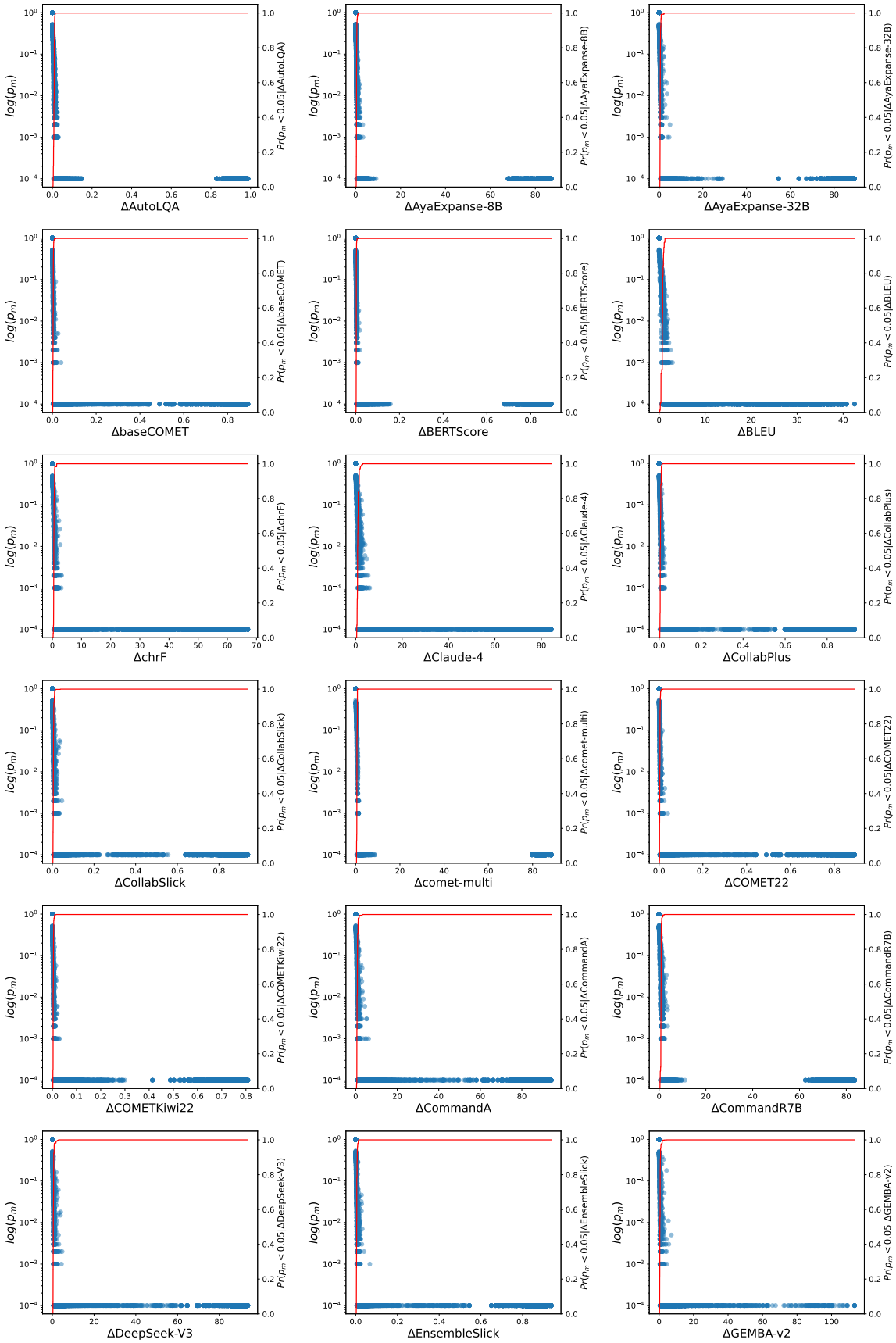

Figure 12: Log p-value of significance test with bootstrap resampling (p_m) on system-level against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \mid \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001. (Part 1/3)

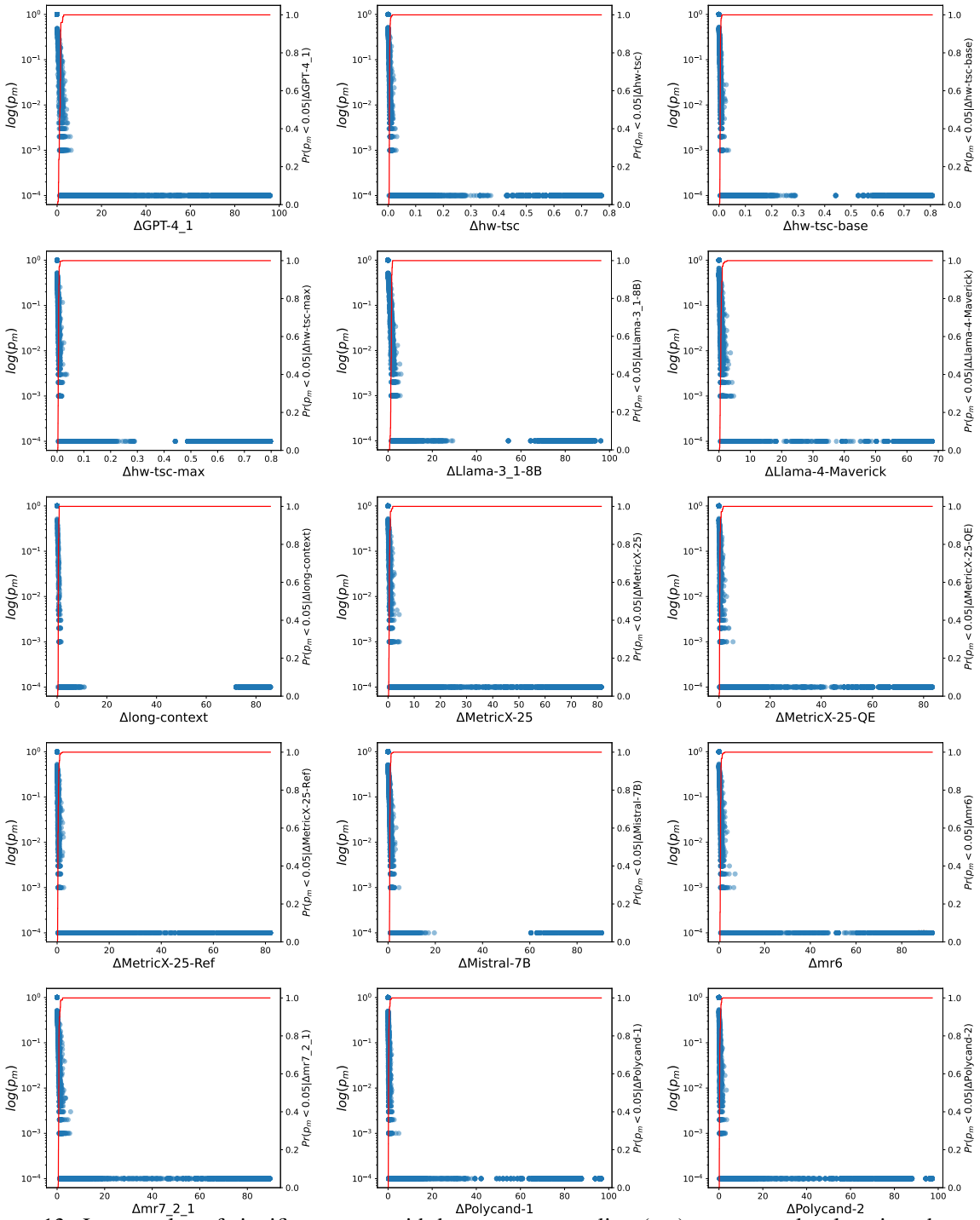

Figure 13: Log p-value of significance test with bootstrap resampling (p_m) on system-level against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \mid \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001. (Part 2/3)

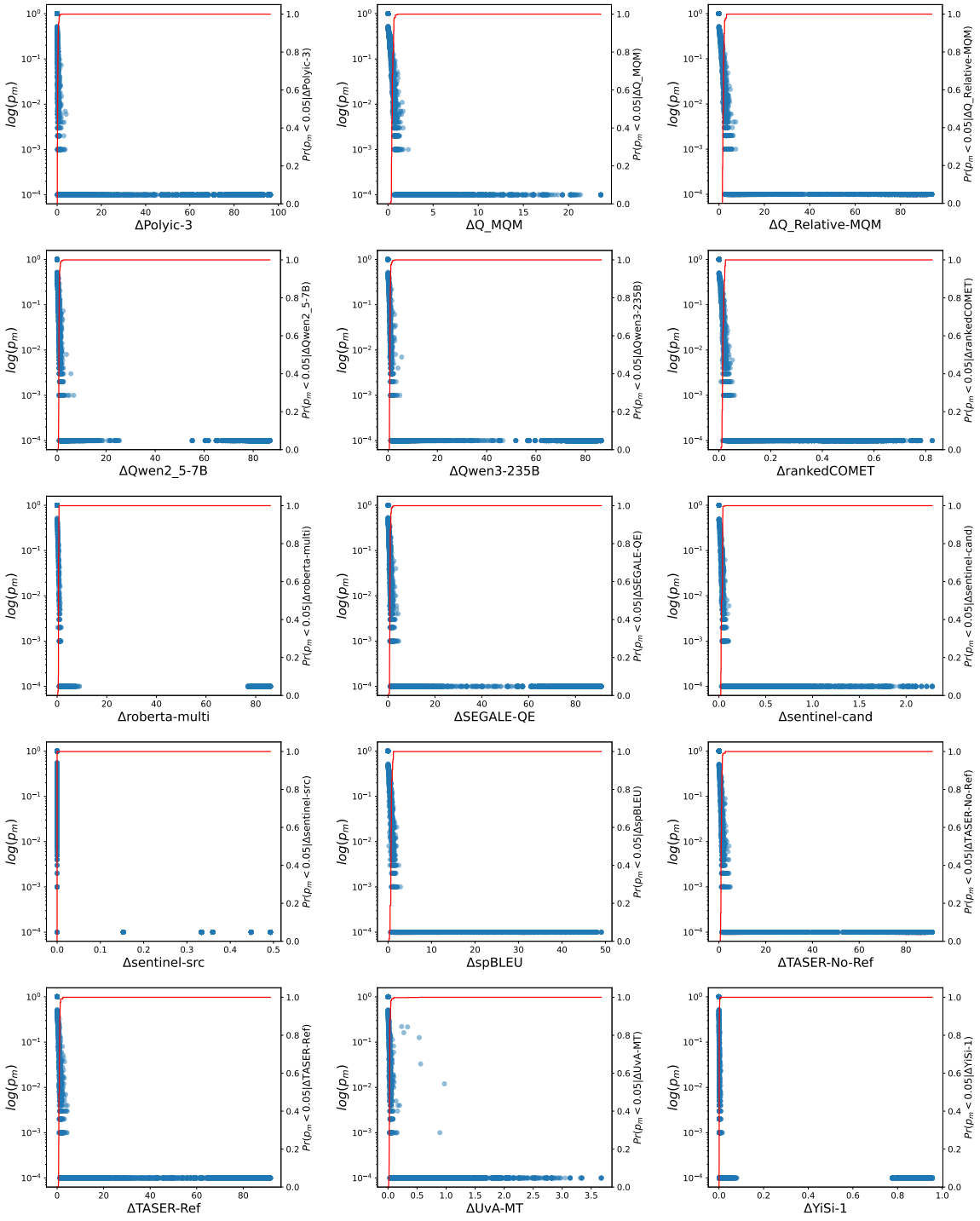

Figure 14: Log p-value of significance test with bootstrap resampling (p_m) on system-level against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \mid \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001. (Part 3/3)

D Task 2 Additional Details

Algorithm 1 Character-Level Error Span F1 Score

function GET_CHAR_F1($L_{hyp}, E_{\text{gold}}, E_{\text{pred}}, \rho$)
▷ L_{hyp}: List of hypothesis lengths
▷ E_{gold}: List of lists of gold error dicts
▷ E_{pred}: List of lists of predicted error dicts
▷ ρ: Partial credit factor (e.g., 0.5)
    $tp \leftarrow 0$
    $\text{total\_gold} \leftarrow 0$
    $\text{total\_pred} \leftarrow 0$
    for $i \in 0 \ldots \text{LENGTH}(L_{hyp}) - 1$ **do**
        $H_{\text{len}} \leftarrow L_{hyp}[i]$
        $G_{\text{maj}}, G_{\text{min}} \leftarrow \text{GET\_COUNTS}(E_{\text{gold}}[i], H_{\text{len}})$
        $P_{\text{maj}}, P_{\text{min}} \leftarrow \text{GET\_COUNTS}(E_{\text{pred}}[i], H_{\text{len}})$
        $\text{total\_gold} \leftarrow \text{total\_gold} + \sum G_{\text{maj}} + \sum G_{\text{min}}$
        $\text{total\_pred} \leftarrow \text{total\_pred} + \sum P_{\text{maj}} + \sum P_{\text{min}}$
        for $j \leftarrow 0$ To $H_{\text{len}} - 1$ **do**
            $c_{\text{g\_maj}} \leftarrow G_{\text{maj}}[j]$
            $c_{\text{p\_maj}} \leftarrow P_{\text{maj}}[j]$
            $c_{\text{g\_min}} \leftarrow G_{\text{min}}[j]$
            $c_{\text{p\_min}} \leftarrow P_{\text{min}}[j]$
▷ Full credit for same severity match at index j
            $tp \leftarrow tp + \min(c_{\text{g\_maj}}, c_{\text{p\_maj}})$
            $tp \leftarrow tp + \min(c_{\text{g\_min}}, c_{\text{p\_min}})$
▷ Partial credit for cross-severity match at index j
            $g_{\text{unmatched}} \leftarrow \max(0, c_{\text{g\_maj}} - c_{\text{p\_maj}}) + \max(0, c_{\text{g\_min}} - c_{\text{p\_min}})$
            $p_{\text{unmatched}} \leftarrow \max(0, c_{\text{p\_maj}} - c_{\text{g\_maj}}) + \max(0, c_{\text{p\_min}} - c_{\text{g\_min}})$
            $tp \leftarrow tp + \min(g_{\text{unmatched}}, p_{\text{unmatched}}) \times \rho$
        end for
    end for
    $P, R, F1 \leftarrow \text{PREC\_REC\_F1}(tp, \text{total\_gold}, \text{total\_pred})$
    return $P, R, F1$
end function

Figure 15: F1 Score by Error Category.

E Full Results for Task 2

We report complete results for all submissions (primary and secondary) in Tables 23 (micro-F1) and 24 (macro-F1) respectively. We also show macro-F1 scores broken down by error category and error ratio in Figure 15 and Figure 16 respectively.

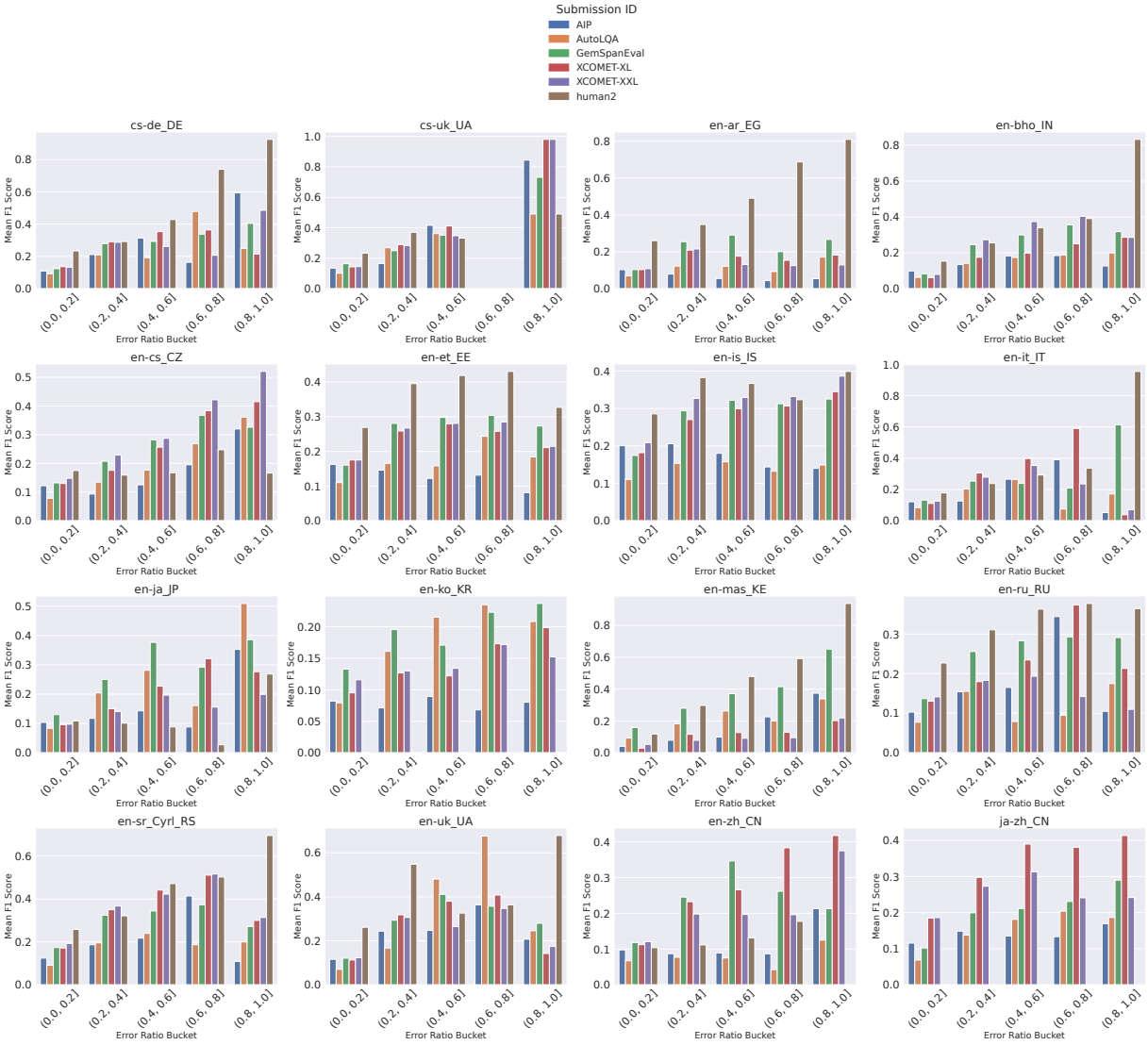

Figure 16: F1 Score by Error Ratio.

Language Pair	Baselines						Primary Submissions									Secondary Submissions									Human2		
	XCOMET-XL			XCOMET-XXL			AutoLQA			AIP			GemSpanEval			AutoLQA			AIP			GemSpanEval					
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CS-DE_DE	24.55	5.15	8.52	25.17	5.22	8.65	17.71	4.02	6.56	11.94	20.37	15.06	28.89	6.02	9.96	15.54	4.58	7.08	14.76	20.31	17.09	31.33	6.66	10.99	30.46	41.08	34.98
CS-UK_UA	25.02	4.00	6.90	28.21	3.56	6.32	20.73	1.93	3.54	13.60	10.16	11.63	35.94	3.58	6.52	17.02	1.99	3.56	18.71	12.03	14.65	37.31	3.96	7.15	27.67	28.95	28.30
EN-AR_EG	11.57	19.53	14.54	8.48	17.92	11.51	11.45	22.95	15.28	2.51	30.41	4.63	19.23	22.18	20.60	10.28	22.65	14.14	2.95	31.76	5.40	21.16	23.50	22.27	79.61	76.37	77.96
EN-BHO_IN	22.87	3.47	6.02	33.46	3.40	6.17	15.55	3.48	5.69	9.38	8.19	8.74	28.40	3.68	6.52	14.91	3.45	5.60	5.54	6.70	6.07	24.57	3.70	6.43	61.31	54.03	57.44
EN-CS_CZ	16.06	6.02	8.76	22.67	6.87	10.55	14.22	4.12	6.39	7.27	15.30	9.85	24.20	6.38	10.10	15.40	3.98	6.32	9.45	15.68	11.79	25.93	6.97	10.99	14.40	24.86	18.24
EN-ET_EE	14.93	16.66	15.75	16.56	17.07	16.81	14.84	11.82	13.16	5.71	20.34	8.92	23.09	12.28	16.04	11.41	10.86	11.13	6.75	22.86	10.42	22.97	12.54	16.23	33.31	32.87	33.09
EN-IS_IS	22.41	27.72	24.78	29.10	28.30	28.70	8.85	19.68	12.21	9.15	35.69	14.57	25.90	19.58	22.30	6.65	17.82	9.69	9.70	34.50	15.14	25.90	20.12	22.64	36.44	40.10	38.18
EN-IT_IT	30.60	4.16	7.33	23.40	5.40	8.77	17.89	2.55	4.47	10.45	13.70	11.86	33.71	5.47	9.41	18.66	2.69	4.71	11.92	15.02	13.29	33.98	5.53	9.51	30.52	30.62	30.57
EN-JA_JP	13.97	3.67	5.81	14.85	3.69	5.92	22.70	2.30	4.18	8.88	11.72	10.10	28.47	3.32	5.94	23.88	2.71	4.86	9.61	10.60	10.08	28.64	3.26	5.85	10.61	13.93	12.04
EN-KO_KR	8.74	14.64	10.95	9.96	17.09	12.58	20.23	7.26	10.69	4.81	25.89	8.12	17.65	10.54	13.20	16.50	6.91	9.75	5.42	28.15	9.08	20.78	11.17	14.53	-	-	-
EN-MAS_KE	10.23	34.84	15.81	11.35	36.31	17.29	15.14	38.95	21.80	27.94	28.65	28.29	35.03	35.67	35.35	14.92	37.87	21.41	4.43	38.08	7.93	35.14	35.72	35.43	94.73	92.14	93.41
EN-RU_RU	16.70	8.59	11.34	17.95	8.48	11.52	13.77	3.84	6.01	8.74	16.77	11.49	28.28	6.49	10.55	12.32	4.00	6.03	9.72	17.06	12.38	29.29	6.39	10.49	25.28	27.56	26.37
EN-SR_CYRL	21.18	21.59	21.38	24.17	21.07	22.52	13.61	15.16	14.35	6.81	27.11	10.88	21.66	15.67	18.18	11.76	15.19	13.26	7.56	26.42	11.75	21.23	17.15	18.97	61.67	58.32	59.95
EN-UK_UA	21.84	2.98	5.25	27.31	3.20	5.73	15.48	1.32	2.43	12.63	6.98	8.99	37.01	2.19	4.13	14.13	1.22	2.24	14.37	6.89	9.32	35.19	2.25	4.22	34.76	39.28	36.88
EN-ZH_CN	22.62	3.80	6.50	19.45	4.14	6.83	13.08	2.92	4.78	7.72	10.43	8.87	30.02	3.37	6.07	11.61	3.19	5.01	9.00	10.40	9.65	30.30	3.47	6.22	11.84	12.82	12.31
JA-ZH_CN	26.89	21.80	24.08	24.07	20.04	21.87	14.83	13.58	14.18	8.47	41.64	14.08	25.35	17.46	20.68	13.11	13.00	13.06	8.50	39.40	13.98	27.67	17.23	21.23	-	-	-
Average	19.39	12.41	12.11	21.01	12.61	12.61	15.63	9.74	9.11	9.75	20.21	11.63	27.68	10.87	13.47	14.26	9.51	8.61	9.27	20.99	11.13	28.21	11.23	13.95	47.04†	48.31†	47.48†

Table 23: Task 2 micro-F1 (%) by language pair for all auto-raters. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

Language Pair	Baselines						Primary Submissions									Secondary Submissions									Human2		
	XCOMET-XL			XCOMET-XXL			AutoLQA			AIP			GemSpanEval			AutoLQA			AIP			GemSpanEval					
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CS-DE_DE	70.91	16.76	13.07	70.82	21.97	16.22	67.13	20.05	13.91	64.51	55.74	36.45	69.95	23.21	17.08	65.66	24.67	16.63	66.03	46.16	31.22	71.35	24.46	19.14	70.56	77.94	64.46
CS-UK_UA	75.93	20.87	17.49	76.62	24.02	19.37	74.50	16.09	11.44	72.11	49.75	37.74	78.98	17.35	15.67	73.35	17.57	11.99	73.87	40.87	32.48	79.37	18.09	16.33	75.86	80.40	67.55
EN-AR_EG	60.85	25.16	10.54	59.32	27.77	10.07	59.48	31.81	11.53	55.38	49.47	18.86	64.28	28.15	12.24	59.01	32.19	10.12	55.70	42.66	14.04	65.07	27.87	12.89	84.95	82.04	79.33
EN-BHO_IN	83.11	27.92	20.00	86.97	13.37	9.88	79.92	10.80	7.14	77.53	32.68	22.44	83.93	8.92	7.01	79.93	8.65	5.46	77.66	21.92	13.40	82.44	10.13	7.71	82.96	83.15	78.86
EN-CS_CZ	71.64	14.84	10.79	73.77	13.70	10.93	70.09	27.18	17.36	68.01	47.65	31.78	73.59	16.22	12.68	70.61	19.95	13.14	69.36	37.75	25.72	74.33	17.54	14.28	70.92	74.52	60.70
EN-ET_EE	66.82	20.71	11.62	67.00	23.50	13.72	65.39	23.96	12.97	62.35	47.05	24.89	70.18	15.65	10.83	64.16	23.35	11.43	63.41	35.98	18.18	70.65	15.69	11.17	76.91	71.50	64.77
EN-IS_IS	60.19	26.54	15.87	63.13	26.79	18.50	52.49	26.35	10.01	53.37	47.42	21.83	62.36	22.23	14.95	51.04	26.97	9.59	54.43	39.01	17.22	63.01	22.05	15.24	68.79	72.75	62.51
EN-IT_IT	71.04	7.13	7.65	66.96	13.38	11.41	64.67	17.00	11.51	60.63	50.35	32.03	68.60	13.82	11.92	64.81	17.29	11.93	61.41	45.63	29.34	68.48	13.89	11.91	65.00	63.88	52.23
EN-JA_JP	77.58	13.98	10.67	77.54	20.40	16.15	79.76	14.14	11.50	76.13	58.14	44.81	80.78	10.04	8.33	79.75	12.77	10.39	77.02	48.80	37.91	80.88	13.94	12.15	76.78	78.98	64.29
EN-KO_KR	45.43	28.05	12.27	45.99	27.36	14.32	50.64	28.32	14.05	42.77	60.19	26.44	51.27	19.28	11.77	48.44	33.92	15.36	43.31	56.78	24.98	53.33	20.63	13.27	-	-	-
EN-MAS_KE	69.29	77.53	49.07	70.33	77.34	49.19	73.73	48.53	27.42	75.57	57.75	36.13	85.55	39.78	31.45	73.76	47.13	26.88	67.09	44.18	15.59	85.56	39.88	31.59	97.13	96.40	96.33
EN-RU_RU	64.66	17.65	13.19	64.86	18.70	13.80	62.52	30.86	19.09	60.07	50.32	30.50	68.39	12.33	10.77	61.55	29.45	18.06	60.85	45.19	27.29	69.13	12.75	11.20	67.48	70.32	58.49
EN-SR_CYRL_RS	65.00	21.77	14.08	66.36	23.09	15.19	60.11	36.45	18.72	57.26	47.01	23.76	66.38	18.92	12.05	59.51	34.75	16.94	58.13	34.44	15.59	66.33	18.95	12.22	72.39	73.76	64.19
EN-UK_UA	79.47	11.05	10.17	80.71	11.26	10.79	77.33	22.65	16.97	76.73	43.18	33.69	82.64	6.07	6.55	77.15	13.71	10.07	77.73	34.16	27.20	82.42	6.19	6.37	80.58	83.82	72.02
EN-ZH_CN	77.07	7.00	6.55	75.95	10.11	8.65	73.00	45.64	32.37	71.78	52.18	38.14	78.53	8.33	7.50	72.27	45.62	32.29	72.51	46.02	33.65	78.76	8.98	8.01	72.96	74.32	59.82
JA-ZH_CN	49.92	28.45	20.17	48.18	31.36	19.78	37.71	50.95	18.90	34.94	66.17	25.83	42.09	56.08	25.74	36.42	49.17	18.48	36.03	61.77	25.03	43.08	56.51	26.76	-	-	-
Average	68.06	22.84	15.20	68.41	24.01	16.12	65.53	28.17	15.93	63.07	50.94	30.33	70.47	19.77	13.53	64.84	27.32	14.92	63.41	42.58	24.30	70.89	20.47	14.39	78.96†	80.24†	71.60†

Table 24: Task 2 macro-F1 (%) by language pair for all auto-raters. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

F Prompts Used for Task 3

```
Translate the following from source_lang to target_lang.  Include only the
translation (without the <>) and nothing else.
>source_text<
```

Table 25: Prompt for Quality-Informed Segment-Level Error Correction task with translating from scratch.

```
Post-edit the following translation from source_lang to target_lang:
>original_translation< given these errors error_spans.  Include only the translation
(without the <>) and nothing else.
>source_text<
```

Table 26: Prompt for Quality-Informed Segment-Level Error Correction task with post-editing existing translation.