

TurCOMET: A Morphologically-Aware Evaluation System for Document-Level Turkish-English Machine Translation

Atahan Uz Hüseyin Emir Akdağ Susan Üsküdarlı

Boğaziçi University

{atahan.uz, huseyin.akdag}@std.bogazici.edu.tr

suzan.uskudarli@bogazici.edu.tr

Abstract

We introduce **TurCOMET**, the first Turkish-specific augmentation of COMET for evaluating translations between Turkish and English. By extending xCOMET-XL with Turkish morphological features and training it on additional Turkish-English data via an additive bottleneck layer, TurCOMET resolves neural evaluation failures caused by sub-word over-segmentation. To validate our metric, we introduce **TurkDoc-MT**, a meticulously curated Turkish-English document-level aligned dataset of 1,322 instances spanning seven domains. Meta-evaluation demonstrates that TurCOMET significantly outperforms the off-the-shelf baseline, yielding absolute Kendall’s τ correlation gains of +0.113 (system-level) and +0.024 (segment-level), with the largest gains in the TR→EN direction. All models, datasets, code, and annotations are publicly released¹ for the benefit of Turkish natural language processing and benchmarking any translation system.

1 Introduction

Large language models (LLMs) have fundamentally advanced document-level machine translation (MT) (Wang et al., 2023; Hendy et al., 2023), requiring evaluation paradigms that move beyond isolated sentence benchmarks and opaque metrics like BLEU. While neural reference-based metrics such as xCOMET (Guerreiro et al., 2024) represent the state-of-the-art, accurately assessing mid-to-low-resource, morphologically rich languages remains a bottleneck. Existing models rely on heavily multilingual XLM-R encoders whose sub-word vocabularies systematically over-segment agglutinative morphology (e.g., in Turkish: *ev*, *evler*, *evlerimizdekilerden*). Consequently, standard metrics

over-rely on surface-level token overlap and heavily under-penalize morphological hallucinations or polarity reversals.

To solve this, we introduce **TurCOMET**, a morphologically-aware variant of xCOMET-XL. Inspired by Glushkova et al. (2023), TurCOMET augments the pooled encoder representation with Turkish-tailored lexical and morphological features via a residual bottleneck layer. Fine-tuned on 35,357 WMT English-Turkish direct assessment (DA) rows, a framework where human raters score translation quality on a continuous scale, TurCOMET acts as an evaluator explicitly sensitized to lemma- and morpheme-level correspondence.

To benchmark current LLMs and validate TurCOMET, we construct **TurkDoc-MT**, a natively document-bounded, multi-domain EN↔TR benchmark spanning seven domains across various linguistic registers (from technical instructions to web crawls). Using this, we evaluate a diverse spectrum of 11 state-of-the-art models, contrasting generalist models (Gemma-3, GPT-OSS, Qwen3.5) with dedicated MT specialists (TranslateGemma). An expert multidimensional quality metrics (MQM) audit confirms that TurCOMET achieves the highest correlation with human judgment, establishing a new evaluation standard.

Our primary contributions are three-fold: (1) **TurCOMET**, an open-source, morphologically-aware metric tailored for Turkish; (2) **TurkDoc-MT**, a highly curated 7-domain document-level benchmark; and (3) a systematic evaluation of 11 LLMs proving that translation specialists vastly outperform massive generalists on morphologically complex tasks.

2 Related Work

Advanced MT Evaluation & Error Detection. As traditional metrics fail to capture nuanced semantic errors, evaluation frameworks have shifted

¹TurCOMET Hugging Face organization, which hosts the TurCOMET checkpoint, the TurkDoc-MT benchmark, the complete MQM annotations, the training data with the exact splits and regression targets, and the training and evaluation pipelines.

toward LLM-as-a-judge mechanisms (Kocmi and Federmann, 2023) and reference-based MQM (Freitag et al., 2021). xCOMET (Guerreiro et al., 2024) forms our baseline, unifying sentence-level DA with fine-grained error span detection.

Combining Neural and Lexical Signals. Neural metrics frequently under-react to specific phenomena, notably morphological perturbations and omissions (Amrhein et al., 2022). To address this, Glushkova et al. (2023) fuse a COMET regressor with lexical-overlap signals via a small bottleneck layer. TurCOMET adopts this architectural bottleneck, but critically substitutes generic signals with a vocabulary explicitly tailored to Turkish agglutinative morphology.

Document-Level MT. While generative LLMs inherently excel at maintaining terminology and register across extended texts (Wang et al., 2023; Hendy et al., 2023), multi-domain benchmarks for evaluating document-level phenomena in agglutinative languages remain scarce. Our TurkDoc-MT corpus directly addresses this gap.

Specialists vs. Generalist LLMs. Massive generalists demonstrate strong zero-shot multilingual capabilities (Gemma Team, 2024; Touvron et al., 2023) but suffer from high inference costs. Consequently, researchers distill MT capabilities into specialized models (Xu et al., 2024; Google Translate Research Team, 2026). These language-specific adaptations leverage optimized tokenization strategies to better handle complex morphology and minimize formal-text hallucinations.

3 TurCOMET Model

While the publicly available xCOMET-XL model represents the state-of-the-art in reference-based machine translation evaluation, it is fundamentally unoptimized for Turkish. This limitation stems from two key factors. First, Turkish represents only a marginal fraction of the WMT DA training signal used to build xCOMET-XL. Second, the agglutinative nature of Turkish—where a single semantic root often accumulates multiple derivational and inflectional suffixes—causes aggressive fragmentation by the underlying XLM-R sub-word tokenizer. This over-segmentation depresses surface overlap scores and dilutes the quality of the pooled sentence representation.

To address the known vulnerabilities of COMET-family metrics in morphologically rich contexts (Amrhein et al., 2022; Glushkova et al.,

2023), we introduce **TurCOMET**. Designed as a Turkish-specific enhancement of xCOMET-XL, TurCOMET introduces a lightweight additive bottleneck layer that integrates Turkish-aware lexical features into the otherwise frozen pre-trained metric.

3.1 Design Principles

The architecture of TurCOMET is governed by three strict design constraints:

- **Preservation of Pre-trained Generalization:** To retain the robust cross-lingual capabilities of the 3.5B-parameter xCOMET-XL checkpoint and avoid prohibitive computational costs, we freeze the base encoder, sentence regression head, and word-level cross-entropy head. Only the new bottleneck layer is fine-tuned.
- **Bidirectional Evaluation:** Rather than fracturing our limited Turkish training data into two unidirectional models, we process both EN→TR and TR→EN translations within a single architecture, conditioned via a binary feature flag.
- **Strict Baseline Adherence:** To guarantee that the specialized metric never underperforms its baseline, the new bottleneck layer is zero-initialized. TurCOMET begins numerically identical to xCOMET-XL and only deviates based on learned morphological signals.

3.2 Architecture

TurCOMET inherits the core components of the Unbabel xCOMETMetric architecture, including the XLM-R-XL encoder (Goyal et al., 2021), layer-wise sparsemax attention pooling, and regression/classification heads. Our sole architectural modification is the introduction of a sentence-level feature bottleneck.

Let $\mathbf{h} \in \mathbb{R}^d$ represent the pooled sentence embedding generated by xCOMET-XL (where $d = 2560$), and let $\mathbf{f} \in \mathbb{R}^k$ denote a vector of Turkish-specific lexical features (where $k = 5$). The bottleneck is a two-layer Multilayer Perceptron (MLP) that projects the concatenated vector $[\mathbf{h}; \mathbf{f}]$ back to the original encoder dimension:

$$\Delta \mathbf{h} = \mathbf{W}_2 \tanh(\mathbf{W}_1 [\mathbf{h}; \mathbf{f}] + \mathbf{b}_1) + \mathbf{b}_2 \quad (1)$$

where $\mathbf{W}_1 \in \mathbb{R}^{b \times (d+k)}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times b}$, with a bottleneck width of $b = 64$. The original pooled embedding is then updated via a residual connection:

$$\mathbf{h}' = \mathbf{h} + \Delta\mathbf{h} \quad (2)$$

This updated representation, $\mathbf{h}'$, is passed seamlessly to the inherited regression head ($\hat{y} = \text{FFN}(\mathbf{h}')$) and the word-level classifier. Crucially, by zero-initializing the output projection $\mathbf{W}_2$ and bias $\mathbf{b}_2$, $\Delta\mathbf{h}$ evaluates to $\mathbf{0}$ at step zero. This ensures the model starts with scores identical to xCOMET-XL, adding a mere $\sim 330\text{K}$ trainable parameters (less than 0.01% of the total network). Figure 1 illustrates the architecture and highlights the point at which the Turkish-specific features are fused into the pipeline.

3.3 Morphologically-Aware Lexical Features

The feature vector $\mathbf{f}$ incorporates five dimensions, each strategically chosen to address a specific vulnerability of neural metrics in agglutinative languages:

- **chrF++(mt, ref):** Character n -gram F-score with word-level n -grams (Popović, 2017). Character overlap is significantly more diagnostic than token overlap for Turkish, as morphological variants share long character prefixes even when sub-words differ.
- **LaBSE cos(src, mt):** The cosine similarity between cross-lingual sentence embeddings (Feng et al., 2022). This provides a reference-free semantic anchor that remains invariant to the hypothesis’s surface morphology.
- **length-z(src, mt; direction):** The standard deviation of the hypothesis-to-source token ratio relative to a direction-specific prior (calculated via Tatoeba/OPUS). High magnitudes flag over-translation or omission.
- **lemma-TER(mt, ref):** Translation Edit Rate (Snover et al., 2006) calculated on lemmatized tokens rather than surface forms. Turkish lemmatization is processed via Zemberek (Akin and Akin, 2007). This feature neutralizes orthographic noise generated by suffix chains, forcing the metric to prioritize core lexical alignment.
- **Direction Flag:** A binary indicator (1.0 for EN→TR, 0.0 for TR→EN) allowing the bottleneck to learn direction-specific feature weightings.

3.4 Training Methodology and Ablation

Dataset and Target Rescaling: The bottleneck is fine-tuned on the full Turkish subset of the WMT DA corpus, encompassing 35,357 rated sentence

pairs from the 2017 and 2018 campaigns. Because raw DA z -scores are unbounded, we apply the BLEURT-20 linear rescaling technique (Sellam et al., 2020) to map the targets cleanly into the $[0, 1]$ range expected by xCOMET-XL. The dataset was stratified by language pair and year into a 90/5/5 split for training, validation, and testing.

Ablation of Word-Level Supervision: We initially experimented with adding a word-level supervision pathway (following Glushkova et al. (2023)) by deriving localized OK/BAD tags from lemma, number, and proper-noun mismatches. However, this ablation actively degraded downstream correlation in our setting. We hypothesize that heuristic tag noise and the resulting sub-word distributional shifts out-weighed the benefits. Consequently, this pathway was omitted, making the bottleneck-only configuration our final model.

Optimization and Inference: We isolated and trained only the 330K bottleneck parameters using AdamW (learning rate of 1.0×10^{-4}), an MSE regression loss, and bf16-mixed precision on a single NVIDIA A100 GPU. On our internal held-out test split, this fine-tuning improved the segment-level Kendall’s τ from 0.368 (baseline xCOMET-XL) to 0.394. During deployment and evaluation, TurCOMET operates using the full prediction blend, combining the sentence regression output with the inherited word-level severity head to ensure direct structural comparability with the baseline.

4 Dataset Construction: TurkDoc-MT

To rigorously evaluate document-level machine translation, we construct **TurkDoc-MT**, a multi-domain, natively document-bounded English-Turkish parallel corpus. Unlike previous benchmarks that artificially flatten discourse by extracting isolated sentences, our pipeline strictly preserves the original document boundaries, ensuring the retention of inter-sentence context, lexical cohesion, and pronoun resolution chains.

4.1 Corpus Selection and Domain Diversity

The benchmark is aggregated from seven distinct domains (data sources), deliberately chosen to span a wide spectrum of linguistic registers (styles and formalities), from highly curated professional translations to extreme-context web crawls.

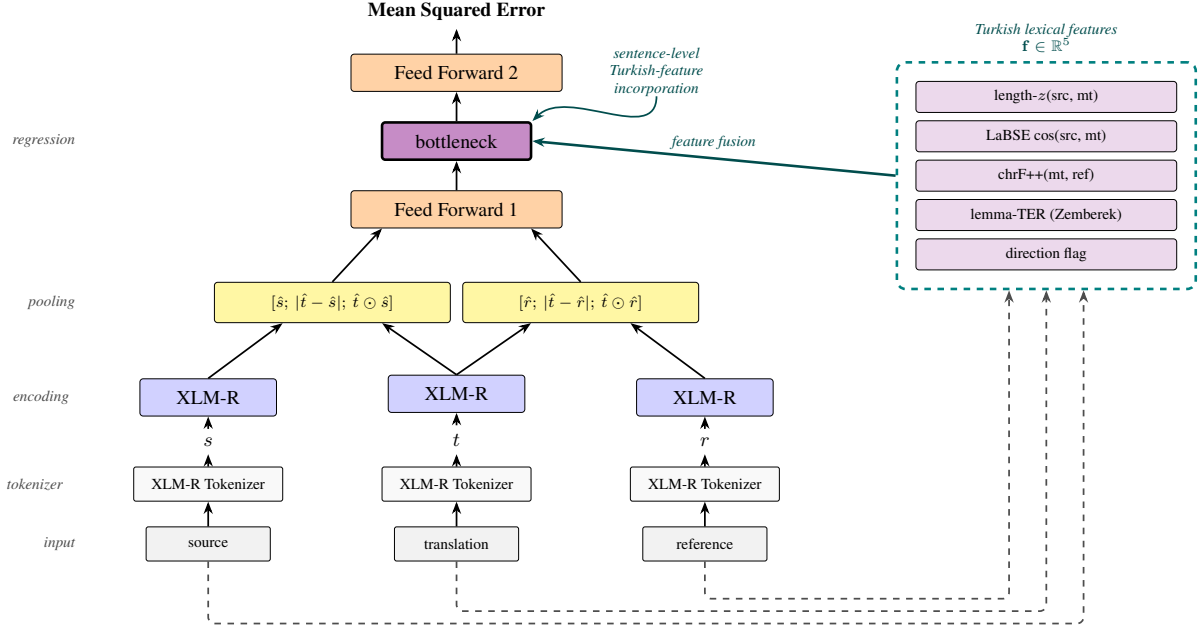


Figure 1: Architecture of **TurCOMET**. The pipeline inherits the frozen xCOMET-XL backbone (blue encoders, yellow pooling, orange feed-forward heads, 3.5B parameters total). Our sole modification is the violet *bottleneck* layer, which receives an additive five-dimensional Turkish-feature vector $\mathbf{f}$ (*length-z*, *LaBSE cosine*, *chrF++*, *lemma-TER*, *direction flag*) computed directly from the source, reference, and translation. The bottleneck is the only trainable component (330K parameters, $< 0.01\%$ of the network) and is zero-initialized, so at step zero the model is numerically identical to xCOMET-XL.

- **High-Quality Baseline (FLORES-200):** We reconstruct full Wikipedia articles from the FLORES-200 dev and devtest splits (NLLB Team et al., 2022), providing a highly curated, grammatically flawless baseline.
- **Journalistic Prose (NTREX-128):** To test standard narrative flow and journalistic idioms, we incorporate the Microsoft NTREX-128 news test set (Federmann et al., 2022), natively aligned by document ID.
- **Educational Dialogue (XQuAD):** Sourced from the Cross-lingual Question Answering Dataset (Artetxe et al., 2020), this domain uniquely tests contextual memory by pairing a formal educational paragraph with informal conversational dialogue.
- **Technical & Scientific (TWiST):** We utilize the Turkish-English Wikipedia Science Translation dataset (Gebeşçe et al., 2025) to test the translation of dense, domain-specific terminology and technical jargon.
- **Extreme Legal Syntax (AYM):** Legal texts represent the ultimate syntactic stress test due to highly nested clauses. We extract official bilingual rulings directly from the Turkish Constitutional Court (Anayasa Mahkemesi) (Zahran, 2025).
- **Spoken Language Domain (TED2020):** To evaluate model adaptability to spoken prose, conversational rhetoric, and presentation transcripts, we integrate the TED2020 talk-level aligned corpus.
- **Internet-Scale Stress Test (DocHPLT):** To evaluate context degradation and output token exhaustion over massive contexts, we sample from the High Performance Language Technologies web crawl (O’Brien et al., 2025), capturing documents exceeding 40,000 words.

4.2 DocHPLT Quality Filtering Mechanism

Raw documents from DocHPLT often suffer from severe alignment noise and structural fragmentation. For instance, vastly different social media posts or forum threads are frequently falsely aligned simply because they share translated boilerplate UI elements (e.g., "Like," "Reply," "Share"). To mitigate this and guarantee a high-fidelity evaluation set, we implement a strict document-level filtering pipeline based on sentence-by-sentence alignment scores.

Formally, let a document D be defined as a set of sentences $S_D = \{s_1, s_2, \dots, s_k\}$. Each sentence s_i possesses an intrinsic cross-lingual alignment score $A(s_i) \in [0, 100]$. We introduce a generalized

bipartite thresholding mechanism parametrized by (m, n) , where $0 \leq m, n \leq 100$. A document D is retained in the final corpus if and only if at least $m\%$ of its constituent sentences meet or exceed an alignment threshold of $n\%$:

$$\frac{|\{s \in S_D \mid A(s) \geq n\}|}{|S_D|} \geq \frac{m}{100} \quad (3)$$

For the construction of the evaluation subset in this work, we enforced a strict isotropic threshold where $m = n = 97$. Consequently, every selected document contains a minimum of 97% of sentences that independently achieved an alignment score of 97% or higher.

4.3 General Dataset Filtering Mechanism

To ensure the high quality of our reference datasets, we implement a two-stage filtering pipeline. Initially, we randomly sample translation pairs and evaluate their quality using OpenAI GPT-5.5² as an automated judge, employing a Scalar Quality Metric (SQM) prompt (see Appendix B). Pairs achieving a score of 3 or higher advance to human expert review, where they are either verified for inclusion or discarded. We repeated this sampling process until we accumulated 200 verified pairs for each dataset, with the exception of NTREX. Because NTREX was limited to an initial pool of 123 translations, our filtering yielded a final evaluation set of 122 pairs.

4.4 Statistical Profile and Context Extremes

TurkDoc-MT consists of 1322 English-Turkish document pairs. A critical feature of **TurkDoc-MT** is its aggressive long-tail length distribution, which mirrors real-world translation demands. Table 1 details the statistical profile of each domain.

Domain	Docs	Mean	90th	99th
FLORES (Wiki)	200	72.8	106.1	124.2
TWiST (Tech)	200	119.9	205.4	363.3
XQuAD (Edu)	200	134.1	190.4	330.5
NTREX (News)	122	342.9	543.6	689.1
AYM (Legal)	200	107.2	193.1	426.3
DocHPLT (Web)	200	225.1	348.1	1,260.0
TED2020 (Speech)	200	2,646.3	3,285.4	4,495.6
Total	1,322	-	-	-

Table 1: Length distribution (in English words) across the seven domains of TurkDoc-MT. The 90th and 99th percentiles illustrate the extreme context variance required for robust LLM evaluation.

²<https://deploymentsafety.openai.com/gpt-5-5>

While human-curated datasets (FLORES, NTREX) natively cluster in the 50 to 500 word range, DocHPLT and especially TED2020 supply the scale needed to test LLM context windows, with 99th percentile boundaries of 1,260 and 4,496 words respectively.

5 Experimental Setup

5.1 Evaluation and Meta-Evaluation Framework

Our experimental design pursues two parallel objectives: the evaluation of LLMs on document translation, and the meta-evaluation of the automated metrics used to score them.

First, for the LLM evaluation, we assessed the bidirectional (EN→TR and TR→EN) translation performance of each model across our full benchmark using three distinct metrics: BLEU, xCOMET-XL, and TurCOMET. Both neural metrics inherit the 512-token context window of their XLM-R backbone, which the source, translation, and reference must share, so longer documents are partitioned into sentence-aligned (source, translation, reference) triples and the document score is the mean over its chunks. The partition is applied identically to xCOMET-XL and TurCOMET and holds the source and reference fixed across all evaluated systems; our released code contains the full implementation.

Second, for the metric meta-evaluation, we assessed the reliability of these automated metrics by comparing them to expert human judgments. To achieve this, we collected fine-grained human annotations on a sampled subset of the **TurkDoc-MT** dataset and calculated the correlation between the metric scores and the human ground truth.

5.2 Models and Prompting Strategy

We evaluated a diverse spectrum of 11 LLMs, ranging from 4B to 120B parameters. As detailed in Table 2, this lineup includes generalist instruction-tuned models (e.g., the Gemma-3), reasoning models (Qwen3.5-9B), and dedicated machine translation specialists (TranslateGemma).

To ensure a controlled and deterministic evaluation environment, we utilized a strict zero-shot prompting strategy with the sampling temperature set to 0. All models received an identical instruction-style prompt adapted from the TranslateGemma technical report (Google Translate Research Team, 2026). This prompt instructed the

model to act as a professional translator between the specified source and target languages and to emit only the translated text without introductory or concluding commentary. The only exceptions were the TranslateGemma models, which utilized their native, specialized prompt formatting. The full prompt templates are provided in Appendix A.

5.3 Human Quality Audit Framework

To validate the automated neural metrics and establish a robust ground truth for English-Turkish document translation, we conducted a rigorous human evaluation study. We recruited four professional, native-speaking Turkish translators to perform a comprehensive quality audit on a stratified subset of the evaluation corpus. To ensure the highest fidelity of evaluation, all four annotators possess formal university-level education in Translation Studies. The annotators contributed on a voluntary, unpaid basis and gave informed consent for their annotations to be released as part of this work.

Crucially, the independent professional specializations of the annotators were intentionally cross-matched with the diverse domain coverage of the **TurkDoc-MT** corpus. Specifically, the cohort includes: (1) an expert in technical and scientific localization, directly mapping to the requirements of the TWiST Tech subset; (2) a translator specializing in literary and creative prose, capturing the unique stylistic and narrative demands of DocHPLT and FLORES; (3) a specialist with extensive experience in formal legal translation, addressing the extreme syntactic complexities of AYM Legal; and (4) an expert in audiovisual and subtitle translation, uniquely aligning with the spoken rhetoric and conversational dialogue registers found in TED2020 and XQuAD.

We sampled 5 translation pairs from each of the 7 datasets, totaling 35 source pairs. Across both translation directions (TR→EN and EN→TR) and all 11 models, this yielded 770 translation outputs for evaluation. Time constraints limited the audit to 550 of these 770 outputs, sampled uniformly at random; the remaining 220 were never rated and are excluded from every analysis. Because the sample is random, the unrated outputs are spread nearly evenly across domains, directions, and systems, and their automatic scores are indistinguishable from those of the rated subset (BLEU 26.0 vs. 25.5, Welch $p = 0.63$; xCOMET-XL 0.735 vs. 0.742, $p = 0.66$; TurCOMET 0.653 vs. 0.654, $p = 0.89$). In total, 2,200 annotations were collected across

the 550 analyzed outputs, so every rated translation carries exactly 4 annotations.

The human evaluation was conducted utilizing the Multidimensional Quality Metrics (MQM) framework, specifically adapting the centralized WMT-Core error typology. Rather than assigning absolute, subjective scalar scores, the expert evaluators operated as fine-grained auditors. They highlighted specific erroneous spans at the sub-word or phrase level, categorized each error into predefined dimensions (Accuracy, Fluency, Terminology, and Style), and assigned a severity weight (Minor, Major, or Neutral). A screenshot of the annotation interface is provided in Appendix C. Finally, to account for individual annotator bias and differing levels of harshness, the severity scores were normalized using z-scores to calculate the final segment-level penalties.

To quantify the reliability of our human evaluation, we computed the inter-annotator agreement (IAA) using Krippendorff’s alpha (α) based on the z-scored penalties. We observed an agreement of $\alpha = 0.660$ at the segment level and a notably higher $\alpha = 0.884$ at the system level. These results indicate that while evaluators exhibit moderate variance when assessing the quality of individual translated segments, there is a strong consensus regarding system-level performance, that is, a model’s average penalty across the segments rated for a given translation direction (e.g., TR → EN), which number 25 per system on average.

Beyond aggregate agreement, we verified that no single annotator, including the one whose professional specialization matched the domain under evaluation, imposed an idiosyncratic standard. Correlating each annotator’s z-scores against the mean of the other three yields $r = 0.51$ to 0.96 across all seven domains, and every pairwise annotator correlation falls between 0.54 and 0.86 . Segment-level agreement is lower for TR→EN than for EN→TR ($\alpha = 0.44$ vs. 0.69), but this is a mechanical consequence of the compressed TR→EN human signal analyzed in Section 6.7 rather than of genuine disagreement: the proportion of segments on which annotators assigned identical scores is in fact higher in that direction (89.6% vs. 64.3%). We note that this cross-matching protocol is language-agnostic and transfers directly to other language pairs.

The models to be evaluated are listed in Table 2.

Model	Params	Source
gpt-oss-120b	120B	openai/gpt-oss-120b
gpt-oss-20b	20B	openai/gpt-oss-20b
gemma-4-31B-it	31B	google/gemma-4-31b-it
gemma-3-27b-it	27B	google/gemma-3-27b-it
gemma-3-12b-it	12B	google/gemma-3-12b-it
gemma-3-4b-it	4B	google/gemma-3-4b-it
translategemma-27b-it	27B	google/translategemma-27b-it
translategemma-12b-it	12B	google/translategemma-12b-it
translategemma-4b-it	4B	google/translategemma-4b-it
Qwen3.5-9B	9B	Qwen/Qwen3.5-9B
Llama-3.1-8B-Instruct	8B	meta-llama/Llama-3.1-8B-Instruct

Table 2: The 11 LLMs evaluated in this work, with parameter counts and Hugging Face source repositories.

6 Results

We evaluate the eleven models introduced in Section 5 along two complementary axes. First, in Section 6.1, we assess the overall translation quality of the models using four independent evaluation signals: a professional human MQM audit, sentence-level BLEU, the public xCOMET-XL checkpoint, and TurCOMET. Second, in Sections 6.2 through 6.6, we conduct a comprehensive meta-evaluation, treating the human MQM audit as ground truth to quantify the reliability and accuracy of each automated metric. Granular per-language and per-dataset breakdowns are provided in the appendix.

6.1 LLM Translation Performance Leaderboards

Table 3 presents the comparative rankings across the four metrics, averaged over both translation directions and sorted by the standardized human MQM penalty (a negated z-score, where higher indicates superior quality across all four annotators).

Model	Humans (neg. z)	BLEU	xCOMET-XL	TurCOMET
gemma-4-31B-it	+0.32	31.3	0.814	0.714
translategemma-27b-it	+0.26	21.4	0.805	0.709
translategemma-12b-it	+0.22	19.0	0.785	0.695
gpt-oss-120b	+0.21	22.5	0.782	0.688
gemma-3-12b-it	+0.21	25.5	0.766	0.675
gemma-3-27b-it	+0.17	28.7	0.782	0.686
gpt-oss-20b	-0.19	22.4	0.749	0.665
Qwen3.5-9B	-0.20	22.6	0.734	0.657
translategemma-4b-it	-0.30	10.1	0.732	0.651
Llama-3.1-8B-Instruct	-0.42	16.2	0.668	0.607
gemma-3-4b-it	-0.43	14.6	0.665	0.606

Table 3: Model leaderboards under four independent evaluation signals (averaged over both translation directions), sorted by the human MQM judges. The human column reports the negated z-score of the MQM penalty across four professional annotators (higher = better, to match the convention of the other metrics). Bold entries indicate the best system per metric.

All four evaluation signals establish **gemma-4-31B-it** as the top-performing system. The dedicated translation specialists, **translategemma-27b-it** and **translategemma-12b-it**, secure the second and third positions under human evaluation and both neural metrics. The metrics agree at the bottom of the table as well, identifying **Llama-3.1-8B-Instruct** and **gemma-3-4b-it** as the weakest systems. However, significant divergences emerge in the middle of the leaderboard, particularly between BLEU and the human judges. For example, **translategemma-27b-it** ranks second according to human evaluators but falls to seventh under BLEU. Conversely, **gemma-3-27b-it** ranks sixth via human evaluation but artificially inflates to second place under BLEU. This discrepancy highlights a fundamental limitation of n -gram overlap metrics: specialized models frequently generate semantically faithful but lexically varied outputs that surface-level string matching disproportionately penalizes. Both xCOMET variants track human rankings more cohesively, motivating the analysis that follows.

6.2 Meta-Evaluation: Metric Agreement with Human Judgments

To rigorously quantify the accuracy with which each automated metric reflects human preference, we computed Pearson (r), Spearman (ρ), and Kendall (τ) correlation coefficients against the z-normalized mean MQM penalties across all four professional annotators ($n = 550$ evaluated translations spanning 22 model-direction systems).

Level	Metric	Pearson	Spearman	Kendall
Segment	BLEU	0.385	0.335	0.256
	xCOMET-XL	0.419	0.350	0.267
	TurCOMET	0.436	0.383	0.291
System	BLEU	0.828	0.852	0.628
	xCOMET-XL	0.950	0.802	0.636
	TurCOMET	0.976	0.913	0.749

Table 4: Correlation between automatic metrics and the z-normalized human MQM penalty, at segment ($n = 550$) and system ($n = 22$) granularity. Higher = better agreement; the sign of the MQM penalty is flipped so that all reported coefficients are positive when metric and humans agree.

As detailed in Table 4, TurCOMET attains the highest correlation with human judgments on the pooled benchmark, under all three coefficients and at both granularities. At the segment level, TurCOMET outperforms xCOMET-XL by approx-

imately +0.024 absolute Kendall τ points and eclipses BLEU by +0.035. This performance gap widens substantially at the system level: TurCOMET achieves a near-perfect Pearson r of 0.976 and a Kendall τ of 0.749, demonstrating that aggregate TurCOMET scores closely mirror the consensus of professional human auditors. These figures pool both directions; Section 6.7 decomposes them and tests significance.

6.3 Pairwise System Ranking Reliability

Beyond absolute correlation, we evaluated pairwise predictive accuracy using the WMT-standard Kendall τ -like formulation (Freitag et al., 2022), $\tau = (C - D)/(C + D)$ over concordant and discordant pairs, which quantifies how often a metric predicts the human preference in a direct A/B comparison.

At the segment level, TurCOMET is concordant with human judgment on 59,056 out of 85,110 pairs ($\tau = 0.388$), outperforming both xCOMET-XL (0.355) and BLEU (0.341). At the system level the margin widens: TurCOMET achieves a τ of 0.749, correctly ordering 202 of 231 system pairs.

6.4 Robustness Across Specific Error Categories

A known vulnerability of reference-based neural metrics is their inconsistent sensitivity to specific localized errors, particularly omissions or numerical hallucinations (Amrhein et al., 2022). Leveraging the fine-grained MQM spans from our human audit, we partitioned the 550 evaluated segments by their dominant error category and computed the Kendall τ correlation within each subset. Table 5 summarizes these findings for categories containing at least six segments.

Dominant error category	n	BLEU	xCOMET-XL	TurCOMET
Accuracy / Mistranslation	103	0.302	0.310	0.267
Other / Non-translation	32	0.017	0.273	0.382
Fluency / Spelling	14	0.456	0.500	0.565
Accuracy / Addition	8	0.286	0.286	0.357
Accuracy / Omission	7	0.250	0.451	0.513
Style / Awkward	7	0.524	0.524	0.619
Fluency / Grammar	6	-0.333	0.600	0.600
Fluency / Inconsistency	8	0.643	-0.286	-0.286

Table 5: Per-error-category Kendall τ between each metric and the human MQM judgement. Translations are partitioned by the category accounting for the largest share of MQM weight within the segment.

TurCOMET achieves the strongest correlation in five of the eight categories and ties in a sixth. It excels in particular at penalizing catastrophic,

safety-critical failures. On segments dominated by *Other / Non-translation* (i.e., severe hallucinations or garbled outputs), TurCOMET achieves a τ of 0.382 compared to a negligible 0.017 for BLEU. It is also the most reliable metric for detecting *Accuracy / Omission* ($\tau = 0.513$) and *Accuracy / Addition* ($\tau = 0.357$), error types that typically bypass standard neural evaluations. Two partitions run the other way: xCOMET-XL leads on *Accuracy / Mistranslation*, the largest at $n = 103$ (0.310 vs. 0.267), and BLEU on the sparse *Fluency / Inconsistency* partition, where surface-level lexical overlap remains an informative signal. Several of these partitions contain fewer than ten segments, so the individual coefficients are indicative rather than conclusive; we report them to show where the metrics diverge, not as a per-category ranking.

6.5 Aggregate ACES-Weighted Performance

To consolidate these fine-grained categorical correlations into a unified, MQM-aligned summary, we adopted the ACES evaluation framework (Glushkova et al., 2023). We calculated the weighted Kendall sum across all error categories, applying the standard MQM severity multipliers (Accuracy $\times 5$, Fluency/Style $\times 1$, Punctuation $\times 0.1$):

$$\text{ACES} = \sum_{c \in \text{categories}} w_c \cdot \tau_c.$$

Metric	ACES-weighted Kendall
TurCOMET	9.09
xCOMET-XL	7.93
BLEU	5.56

Table 6: ACES-style (Glushkova et al., 2023) weighted Kendall τ sum across MQM categories. Weights follow the MQM severity rubric (Accuracy categories $\times 5$, Fluency/Style $\times 1$, Punctuation $\times 0.1$).

As shown in Table 6, TurCOMET achieves the highest overall ACES-weighted score (9.09), numerically confirming its consistent robustness in heavily penalizing high-severity accuracy errors.

6.6 Metric Sensitivity by Error Severity

To evaluate metric behavior across different error magnitudes, we partitioned segments by maximum MQM error severity into major-heavy (171 segments) and minor-heavy (17 segments) subsets. For major errors, all metrics correlate moderately with

human judgments ($r > 0.41$). However, for subtle minor errors, BLEU’s correlation deteriorates to random noise ($r \approx 0$). In contrast, TurCOMET ($r = 0.428$) and xCOMET-XL ($r = 0.470$) maintain strong discriminative power, confirming TurCOMET as a robust proxy for human evaluation across all error severities.

6.7 Directional Robustness and Statistical Significance

The correlations reported so far pool both translation directions, which place very different demands on a morphologically aware metric. Table 7 separates them and tests whether the TurCOMET–xCOMET-XL gap exceeds sampling noise.

Direction	BLEU	xCOMET-XL	TurCOMET
EN→TR	0.264 / 0.816	0.329 / 0.970	0.323 / 0.978
TR→EN	0.008 / 0.751	0.077 / 0.632	0.127 / 0.888

Table 7: Per-direction agreement with the z -normalized human MQM penalty, pooled over all seven domains. Cells report segment-level Kendall τ / system-level Pearson r ; the best metric is bolded separately per granularity. Computed from the per-segment scores of the 550 rated translations, these coefficients cannot be re-derived from Table 10, whose columns average over the full benchmark.

Significance. We test the segment-level difference between TurCOMET and xCOMET-XL with a paired bootstrap over segments ($B = 5,000$) and the system-level difference with Williams’ test. Across all 550 rated segments TurCOMET improves Kendall τ by $\Delta\tau = +0.0247$ (95% CI $[+0.004, +0.049]$, $p = 0.022$); the gain is carried almost entirely by TR→EN ($\Delta\tau = +0.049$, $p < 0.001$; Williams’ $p = 0.048$ at system level). In EN→TR the two neural metrics are statistically indistinguishable ($p = 0.71$ segment, $p = 0.56$ system).

Where the gain originates. Pooled over both directions TurCOMET is the strongest metric, yet the improvement concentrates in TR→EN. This is a decomposition, not a contradiction: in EN→TR the baseline already sits near its system-level ceiling ($r = 0.970$), leaving little headroom, and TurCOMET matches it ($r = 0.978$). We attribute the TR→EN gain to the direction flag and the LaBSE feature, which anchors the English hypothesis to the Turkish source rather than the reference; absent a per-feature ablation this remains an attribution rather than a demonstration.

Compression of the TR→EN human signal.

The low absolute TR→EN coefficients reflect the human ground truth, not the metrics: all four annotators judged 82.9% of rated TR→EN segments error-free against 48.1% for EN→TR, leaving 68.7% of TR→EN segment pairs tied (23.6% for EN→TR). All coefficients we report use the tie-corrected τ_b ; the WMT-style τ that discards tied pairs widens TurCOMET’s TR→EN margin to 0.227, against 0.138 for xCOMET-XL and 0.014 for BLEU, and moves EN→TR by at most 0.006. The advantage therefore sits on precisely those pairs where annotators did perceive a difference, and at system level, where averaging restores the signal, TurCOMET reaches $r = 0.888$.

Genre dependence. The per-domain breakdown in Appendix F shows BLEU or xCOMET-XL ahead of TurCOMET on individual domains. Each cell rests on only 33–46 rated translations (CI $\approx \pm 0.32$ on a correlation), so the per-domain winners cannot be separated statistically; we read that table as evidence of genre variability, not a per-genre ranking.

7 Conclusion

We introduced **TurCOMET**, a morphologically-aware metric that addresses the architectural shortcomings of standard neural evaluators on agglutinative languages. By injecting Turkish-specific lexical signals and lemma-level edit rates into a frozen cross-lingual encoder via an additive bottleneck layer, TurCOMET mitigates sub-word over-segmentation and achieves state-of-the-art correlation with expert human judgment. To validate it, we released **TurkDoc-MT**, a manually curated 1,322-document benchmark spanning seven domains. Our MQM audit of 11 modern LLMs establishes a statistically significant improvement over xCOMET-XL overall and in TR→EN, with parity in EN→TR, and shows that dedicated translation specialists outperform massive generalist architectures on morphologically complex tasks. Beyond Turkish, TurCOMET is best understood as a lightweight language-adaptation recipe that transfers to any language with a lemmatizer and rated translation data. Validating it on typologically similar languages is our primary line of future work; all artifacts are publicly released.

Limitations

While the TurkDoc-MT benchmark provides a robust multi-domain evaluation, we note the following limitations in our current study:

- **Annotation Density and Coverage:** Due to the high cost of expert-level MQM annotation, our human evaluation covers 550 of the 770 generated translations, each rated by all four annotators. While sufficient for aggregate correlation analysis, four ratings per segment may underrepresent annotator disagreement on highly ambiguous or subjective texts.
- **Chunked Metric Scoring:** The 512-token window of the XLM-R backbone prevents xCOMET-XL and TurCOMET from consuming a long document in a single forward pass, so both score each document as a sequence of independently scored, sentence-aligned chunks. Discourse phenomena spanning chunk boundaries therefore fall outside the neural metrics’ view even though the benchmark itself preserves them, and long-context metric design remains open future work.
- **Inherited Error Spans:** A notable feature of xCOMET is its ability to highlight erroneous spans within translations. TurCOMET freezes the word-level classification head, and its bottleneck alters only the pooled representation feeding the sentence-level regression head, so its span predictions are identical to those of xCOMET-XL by construction. We neither evaluated nor improved span detection, focusing instead on aligning the overall numeric quality scores with human preferences; the MQM spans we release make this a natural next step.
- **Directional and Genre Dependence:** As quantified in Section 6.7, TurCOMET’s advantage over xCOMET-XL is significant overall and in TR→EN, but not in EN→TR, where the baseline already sits near its system-level ceiling. The per-domain cells of Appendix F are too small ($n = 33\text{--}46$) to rank the metrics by genre. TurCOMET should therefore be read as the strongest metric on this benchmark in aggregate, not as uniformly superior in every direction and domain.
- **Single-Language Validation:** We validate the approach on Turkish alone. The design is intended as a language-adaptation recipe rather than a Turkish-only artifact: only the 330K-parameter bottleneck ($< 0.01\%$ of the network) is trained, and of the five input signals only lemma-TER requires language-specific tool-

ing (here, Zemberek). Adapting the method to another morphologically rich language requires a lemmatizer and rated translation data, with MQM annotation needed only for meta-evaluation. Whether a single multilingual adaptation could match a set of per-language bottlenecks is an open question we cannot answer from Turkish alone.

- **Limited Model Coverage:** Due to constraints on computational resources, the range of evaluated models is restricted. Consequently, we were unable to include the latest frontier LLMs, such as DeepSeek-V4-Pro³, in our current benchmarking efforts.

Acknowledgments

We extend our deepest gratitude to the four professional translators who volunteered their time, domain expertise, and effort to conduct the comprehensive Multidimensional Quality Metrics (MQM) audit.

References

- Mehmet Dündar Akın and Ahmet Afşin Akın. 2007. Türk dilleri için açık kaynaklı doğal dil işleme kütüphanesi: ZEMBEREK. *Elektrik Mühendisliği*, 431:38–44. Software: <https://github.com/ahmetaa/zemberek-nlp>.
- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513. Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637. Association for Computational Linguistics.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics.

³<https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro>

- 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, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68. Association for Computational Linguistics.
- Ali Gebeşçe, Gözde Gül Şahin, and Ege Uğur Amasya. 2025. [TWiST: Turkish-English Wikipedia & Thesis STEM Terminology Dataset](#). Version 1.
- Gemma Team. 2024. [Gemma: Open models based on Gemini research and technology](#). *arXiv preprint arXiv:2403.08295*.
- Taisiya Glushkova, Chrysoula Zerva, and André F. T. Martins. 2023. [BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 47–58. European Association for Machine Translation.
- Google Translate Research Team. 2026. [TranslateGemma technical report](#). *Preprint*, arXiv:2601.09012.
- Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. [Larger-scale transformers for multilingual masked language modeling](#). In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. [How good are GPT models at machine translation? a comprehensive evaluation](#). *arXiv preprint arXiv:2302.09210*.
- 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.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and Jean Maillard. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Dayyán O’Brien, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. [DocHPLT: A massively multilingual document-level translation dataset](#). *arXiv preprint arXiv:2508.13079*.
- 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.
- 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.
- 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.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661. Association for Computational Linguistics.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). *Preprint*, arXiv:2309.11674.
- Abdelrahman Zahran. 2025. [Bidirectional machine translation dataset \[EN-TR\]](#). <https://www.kaggle.com/datasets/abdelrahmanzahran/bidirectional-machine-translation-dataset-en-tr>. Version 1, published 2025-05-17. MIT License. Accessed: 2026-05-08.

A Prompt Template

All models were evaluated using the following instruction-style prompt, adapted from the TranslateGemma technical report (Google Translate Research Team, 2026). The placeholders {source_lang}, {src_lang_code}, {target_lang}, and {tgt_lang_code} are populated per direction (e.g., English/en, Turkish/tr), and [TEXT] is replaced with the source document.

```
You are a professional {source_lang} ({src_lang_code}) to {target_lang} ({tgt_lang_code}) translator. Your goal is to accurately convey the meaning and nuances of the original {source_lang} text while adhering to {target_lang} grammar, vocabulary, and cultural sensitivities. Produce only the {target_lang} translation, without any additional explanations or commentary. Please translate the following {source_lang} text into {target_lang}: [TEXT]
```

B SQM Judge Prompt for Dataset Filtering

The following Scalar Quality Metric (SQM) prompt was used with GPT-5.5 as a judge during the general dataset filtering pipeline described in Section 5. The placeholders {en} and {tr} are populated with the English source and Turkish translation, respectively.

```
You are a professional translation quality evaluator performing Direct Assessment (WMT-style). Evaluate the following English-to-Turkish translation. Consider: - Adequacy: does the Turkish preserve the meaning of the English source? - Fluency: is the Turkish grammatical, natural, and well-formed? - Accuracy: are there mistranslations, omissions, additions, or hallucinations? Score as an INTEGER from 0 to 6 (no decimals):
0: "0 - Nonsense / No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.",
1: "1 - Between 0 and 2",
2: "2 - Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.",
3: "3 - Between 2 and 4",
4: "4 - Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.",
5: "5 - Between 4 and 6",
6: "6 - Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.",
If either side is empty, garbage, in the wrong language, or boilerplate (cookie notices, menus, navigation), score very low.
English source: {en}
Turkish translation: {tr}
Respond with ONLY a JSON object in this exact format, no other text: {"score": <integer between 0 and 6>}
```

C Annotation Interface

Figure 2 shows a screenshot of the MQM-based human annotation interface used by the professional translators to highlight erroneous spans, assign error dimensions (Accuracy, Fluency, Terminology, Style), and select a severity weight (Minor, Major, Neutral).

English (source)	Turkish (model translation)
However, the restriction to be imposed on the freedoms of expression and the press must be in conformity with the principles of “being necessary in a democratic society” and “being proportionate”, which are among the general restriction criterion stipulated in Article 13 of the Constitution. The principle of being necessary in a democratic society must be interpreted based on pluralism, tolerance and open-mindedness. The principle of proportionality reflects the relationship between the aim of restriction and the means	Ancak, ifade ve basın özgürlüğüne getirilecek kısıtlamalar, Anayasa'nın 13. maddesinde öngörülen genel kısıtlama koşulları arasında yer alan "demokratik bir toplumda gerekli" ve "orantılı" ilkelerine uygun olmalıdır. "Demokratik bir toplumda gerekli" ilkesi, çoğulculuk, hoşgörü ve açık
	Selected span Demokratik bir toplumda gerekli" Category Accuracy / Mistranslation Severity Major (weight 5) Note (optional) Cancel Add error

Figure 2: Screenshot of the human evaluation interface used for the MQM-based quality audit.

D TurCOMET Training Configuration

This appendix details the full training configuration used to produce the final TurCOMET checkpoint described in Section 3.4. The reported run corresponds to the recommended setting (bottleneck-only fine-tuning over plain machine-translation inputs, 6 epochs), which strictly outperforms the xCOMET-XL baseline on every reported correlation metric in both regression-only and full prediction-blend inference modes on our internal held-out split.

Hardware and Framework. All training was performed on a single NVIDIA A100-SXM4-80GB (80 GB VRAM) instance with 167 GB of system memory, using PyTorch 2.10.0 (CUDA 12.8) and PyTorch Lightning. The backbone is the public Unbabel/XCOMET-XL checkpoint, loaded via XCOMETMetric (unbabel-comet 2.2.7). The full forward pass uses bf16-mixed automatic mixed precision.

Trainable Parameters. Of the 3.49 B total model parameters, only the 330 K parameters of the feature bottleneck are trainable; the XLM-RoBERTa-XL encoder, the layer-wise sparsemax attention, the sentence-level regression head (estimator), and the word-level severity head (hidden2tag) are all kept frozen for the entirety of the run. The bottleneck consists of $\text{Linear}(2565 \rightarrow 64) \rightarrow \text{Tanh} \rightarrow \text{Linear}(64 \rightarrow 2560)$, with the final Linear zero-initialized so that the model begins numerically identical to xCOMET-XL.

Data Split. The full Turkish subset of the WMT DA corpus (35,357 segments combining the 2017 and 2018 EN $\leftrightarrow$ TR campaigns) is stratified by language-pair and year and split 90/5/5 into 31,821 training, 1,768 validation, and 1,768 test segments. Sentence-level regression targets are the score_minmax column, i.e. the BLEURT-20 linear rescaling of the raw DA z -scores into $[0, 1]$.

Hyperparameters. Table 8 lists the complete set of hyperparameters used for the final run.

Training Schedule. Training proceeded for the full six epochs (5,969 optimization steps); the early-stopping criterion (patience 3 on val_kendall) was never triggered, as a new best validation score was achieved at the final evaluation. Peak GPU memory at steady state was 30–40 GB, well within the 80 GB ceiling of the A100. The best checkpoint, epoch=5-step=5969-val_kendall=0.364.ckpt, reaches a validation Kendall’s τ of 0.3638 and is the checkpoint released as TurCOMET.

Hyperparameter	Value
<i>Architecture</i>	
Backbone	Unbabel/XCOMET-XL
Encoder	XLm-RoBERTa-XL (3.5 B)
Feature dim	5
Bottleneck dim	64
Bottleneck init	zero (final Linear)
Sentence layer	mix (sparsemax)
Word layer	36
Hidden sizes (FFN)	[2560, 1280]
Activation	Tanh
Dropout	0.1
<i>Optimization</i>	
Optimizer	AdamW
Bottleneck LR	1.0×10^{-4}
Encoder LR	5.0×10^{-8} (frozen)
FFN head LR	5.0×10^{-7} (frozen)
Layerwise decay	0.983
Sentence loss	MSE
Word-level loss weight (λ)	0
Gradient clip	1.0 (norm)
Precision	bf16-mixed
<i>Schedule</i>	
Batch size	16
Accumulate grad batches	2 (effective 32)
Max epochs	6
Min epochs	1
Val check interval	0.5 (twice per epoch)
Early-stopping monitor	val_kendall (mode: max)
Early-stopping patience	3
Seed	12
<i>Inputs</i>	
MT field used	plain mt (no inline tags)
Score column	score_minmax (BLEURT-20 rescaled)

Table 8: Final training configuration for TurCOMET. The full run completes in approximately 100 minutes on an NVIDIA A100-80GB.

E Representative Dataset Samples

To illustrate the extreme linguistic diversity and domain variance of the **TurkDoc-MT** benchmark, Table 9 presents representative document-level samples from all seven subsets of our evaluation corpus. These examples highlight the variation in sentence length, register, and the agglutinative density of the Turkish translations.

English Source Document	Turkish Reference Translation
Domain: FLORES-200 (Encyclopedic / Instructional)	
Internal poisoning may not be immediately apparent. Symptoms, such as vomiting are sufficiently general that an immediate diagnosis cannot be made.	İç zehirlenme hemen ortaya çıkmayabilir. Kusma gibi belirtiler hemen tanı konulamayacak kadar geneldir.
Domain: TWiST (Technical & Scientific)	
The application of quantum mechanics to physical objects such as the electromagnetic field, which are extended in space and time, is known as quantum field theory.	Kuantum mekaniğinin uzay ve zamanda genişleyen elektromanyetik alan gibi fiziksel objelere uygulanması kuantum alan teorisi olarak bilinir.
Domain: XQuAD (Educational Dialogue & QA)	
The Rhine-Meuse Delta begins near Millingen aan de Rijn. Question: Where does the Delta begin? Answer: near Millingen aan de Rijn	Rhine-Meuse Deltası, Millingen aan de Rijn civarında başlar. Soru: Delta nerede başlar? Cevap: Millingen aan de Rijn civarında
Domain: NTREX-128 (Journalistic Prose)	
Almanac: The inventor of the Geiger Counter. September 30, 1882, the future physicist Johannes Wilhelm "Hans" Geiger was born in Germany.	Almanak: Geiger Sayacının mucidi. 30 Eylül 1882, geleceğin fizikçisi Johannes Wilhelm "Hans" Geiger'in Almanya'da doğduğu gün.
Domain: AYM (Extreme Legal Syntax)	
For the reasons explained above and in the light of the relevant case-law of the Constitutional Court and the ECHR, I am of the opinion that while the alleged violations satisfy the admissibility criteria, finding of a violation is not appropriate.	Yukarıda zikredilen gerekçeler doğrultusunda ve Anayasa Mahkemesi ile AIHM'in benzer kararları ışığında; başvuruyu kabul edilebilirlik kriterleri açısından yeterli saymakla birlikte, hak ihlaline karar verilmesinin uygun olmadığını düşünüyorum.
Domain: TED2020 (Spoken Conversational Rhetoric)	
I'd like to talk today about how we can change our brains and our society. Meet Joe. Joe's 32 years old and a murderer.	Bugün beynimizi ve toplumuzu nasıl değiştirebileceğimize dair konuşmak istiyorum. Joe ile tanışın. Joe 32 yaşında bir katil.
Domain: DocHPLT (Extreme-Context Web Crawl)	
Caribbean Pirate Girl's Journey Elizabeth was on her way to America when she saw the pirate ship! Help her find the treasure map before she gets into trouble! ARRR!	Karayipli Korsan Kızın Güncesi Korsan gemisini gördüğünde Elizabeth Amerika yolundaydı! Başı belaya girmeden önce hazine haritasının yerini bildiği yalanını söyledi! HIRRR!

Table 9: Bilingual document samples from the TurkDoc-MT evaluation dataset. The examples highlight the variation in register (ranging from objective encyclopedic entries to highly idiomatic web content), context length, and the agglutinative complexity of the Turkish translations.

F Per-Direction and Per-Dataset Results

The main text reports model leaderboards and metric-vs-human correlations averaged over both translation directions and all seven evaluation domains. This appendix provides the full breakdown along both axes. Table 10 reports the four-metric leaderboard separately for EN→TR and TR→EN. Tables 11 and 12 report the per-dataset leaderboards for each translation direction, respectively. Table 13 reports the segment-level Pearson, Spearman, and Kendall correlations between each automatic metric and the z -normalised mean MQM penalty, broken down by dataset and direction.

Two observations are worth emphasizing. First, the TR→EN direction shows substantially less variance in human judgement than EN→TR; many cells in the TR→EN tables are tied at the +0.42 floor, corresponding to a unanimous zero-error annotation. This compression of the human signal naturally bounds the per-cell correlations achievable for that direction, as reflected in the correlation table. Second, while TurCOMET is the strongest metric on aggregate, the per-dataset and per-direction breakdown reveals that on several individual cells, either BLEU (e.g., TED2020 EN→TR, NTREX EN→TR) or xCOMET-XL (e.g., AYM EN→TR, DocHPLT EN→TR) outperforms TurCOMET. These results suggest that the best-performing metric remains genre-dependent, even when the overall ranking is clear. Both observations are quantified in Section 6.7, which reports the per-direction correlations pooled over domains together with paired bootstrap and Williams’ significance tests. We stress that each individual cell of Table 13 rests on only 33 to 46 rated translations, corresponding to a confidence interval of roughly ± 0.32 on a correlation coefficient; the per-domain winners are therefore not statistically separable, and the bolding in that table marks the numerically highest value only.

Model	Humans (neg. z)	BLEU	xCOMET-XL	TurCOMET
Direction: EN→TR				
translategemma-12b	+0.27	18.8	0.776	0.684
gemma-4-31B	+0.24	24.8	0.796	0.698
translategemma-27b	+0.19	19.3	0.789	0.693
gemma-3-12b	+0.12	18.5	0.730	0.644
gpt-oss-120b	+0.05	17.8	0.753	0.662
gemma-3-27b	-0.02	22.5	0.751	0.659
translategemma-4b	-0.60	6.1	0.676	0.606
gpt-oss-20b	-0.68	14.6	0.703	0.628
Qwen3.5-9B	-0.72	14.1	0.664	0.604
gemma-3-4b	-0.91	8.0	0.608	0.561
Llama-3.1-8B	-1.15	7.0	0.564	0.530
Direction: TR→EN				
gemma-4-31B	+0.41	37.7	0.832	0.730
gpt-oss-120b	+0.37	27.3	0.811	0.715
gemma-3-27b	+0.36	35.0	0.813	0.714
translategemma-27b	+0.33	23.5	0.821	0.724
Qwen3.5-9B	+0.33	31.1	0.803	0.711
Llama-3.1-8B	+0.32	25.4	0.772	0.685
gemma-3-12b	+0.30	32.4	0.802	0.707
gpt-oss-20b	+0.29	30.2	0.795	0.702
translategemma-12b	+0.17	20.5	0.795	0.706
gemma-3-4b	+0.05	21.2	0.721	0.652
translategemma-4b	+0.00	14.0	0.787	0.697

Table 10: Per-direction model leaderboards under the four evaluation signals, reported separately for EN→TR and TR→EN (the main text reports the average of both). The Humans column reports the negated z -score of the mean MQM penalty across the four professional annotators (higher = better). Bold entries indicate the best system per column.

Dataset	Model	Hum	BLEU	xCOM	TurC	Dataset	Model	Hum	BLEU	xCOM	TurC
FLORES	gemma-4-31B	+0.03	28.2	0.876	0.744	NTREX	gemma-4-31B	-0.15	25.3	0.644	0.596
	tgemma-27b	+0.11	24.4	0.874	0.741		tgemma-27b	-0.10	22.0	0.627	0.594
	tgemma-12b	+0.35	21.1	0.866	0.736		tgemma-12b	-0.03	20.0	0.619	0.588
	gpt-oss-120b	-0.61	25.4	0.829	0.703		gpt-oss-120b	-0.12	17.7	0.545	0.527
	gemma-3-12b	+0.42	24.9	0.800	0.680		gemma-3-12b	-0.11	20.8	0.520	0.520
	gemma-3-27b	-0.05	27.5	0.820	0.694		gemma-3-27b	-0.56	22.2	0.560	0.537
	gpt-oss-20b	-0.16	22.3	0.752	0.654		gpt-oss-20b	-1.40	15.9	0.507	0.514
	Qwen3.5-9B	-0.77	17.2	0.782	0.670		Qwen3.5-9B	-2.55	14.8	0.434	0.471
	tgemma-4b	-0.78	19.3	0.776	0.663		tgemma-4b	-0.78	16.3	0.449	0.481
	Llama-3.1-8B	-0.65	17.7	0.636	0.560		Llama-3.1-8B	-2.72	11.5	0.332	0.426
	gemma-3-4b	-0.18	19.0	0.688	0.600		gemma-3-4b	-0.49	13.7	0.351	0.432
TWiST	gemma-4-31B	+0.28	30.6	0.826	0.706	AYM	gemma-4-31B	+0.33	31.9	0.898	0.774
	tgemma-27b	+0.11	27.6	0.816	0.696		tgemma-27b	+0.42	24.5	0.883	0.762
	tgemma-12b	+0.18	24.6	0.816	0.696		tgemma-12b	+0.42	21.3	0.871	0.753
	gpt-oss-120b	+0.19	27.4	0.794	0.676		gpt-oss-120b	+0.42	25.5	0.865	0.745
	gemma-3-12b	-0.66	28.6	0.780	0.664		gemma-3-12b	+0.42	25.2	0.851	0.731
	gemma-3-27b	-0.07	30.6	0.783	0.665		gemma-3-27b	-0.01	28.1	0.865	0.744
	gpt-oss-20b	-0.63	27.0	0.755	0.650		gpt-oss-20b	-0.36	22.1	0.814	0.703
	Qwen3.5-9B	-0.35	21.2	0.759	0.649		Qwen3.5-9B	+0.28	20.1	0.807	0.698
	tgemma-4b	-0.41	23.0	0.743	0.634		tgemma-4b	+0.33	17.9	0.759	0.662
	Llama-3.1-8B	-1.07	19.7	0.612	0.538		Llama-3.1-8B	+0.02	17.3	0.670	0.592
	gemma-3-4b	-0.68	23.8	0.694	0.596		gemma-3-4b	+0.12	19.2	0.683	0.607
XQuAD	gemma-4-31B	+0.30	32.4	0.710	0.646	DocHPLT	gemma-4-31B	+0.42	33.8	0.794	0.700
	tgemma-27b	+0.13	29.1	0.707	0.647		tgemma-27b	+0.23	25.4	0.792	0.697
	tgemma-12b	+0.13	27.1	0.694	0.643		tgemma-12b	+0.32	26.9	0.780	0.688
	gpt-oss-120b	-0.06	29.4	0.679	0.619		gpt-oss-120b	+0.19	27.8	0.765	0.672
	gemma-3-12b	+0.04	29.4	0.658	0.602		gemma-3-12b	+0.42	30.2	0.749	0.661
	gemma-3-27b	-0.21	32.1	0.676	0.616		gemma-3-27b	+0.15	29.1	0.749	0.661
	gpt-oss-20b	-0.58	29.9	0.636	0.593		gpt-oss-20b	-0.09	26.0	0.727	0.649
	Qwen3.5-9B	-0.04	19.2	0.637	0.588		Qwen3.5-9B	-0.02	20.0	0.714	0.636
	tgemma-4b	+0.02	22.9	0.612	0.570		tgemma-4b	-0.12	17.1	0.703	0.632
	Llama-3.1-8B	-0.71	21.7	0.508	0.508		Llama-3.1-8B	-0.47	18.8	0.588	0.545
	gemma-3-4b	+0.42	24.0	0.568	0.541		gemma-3-4b	-0.51	17.3	0.649	0.589
TED2020	gemma-4-31B	+0.42	22.6	0.769	0.683						
	tgemma-27b	+0.42	16.8	0.762	0.677						
	tgemma-12b	+0.42	16.7	0.726	0.649						
	gpt-oss-120b	+0.15	14.7	0.718	0.639						
	gemma-3-12b	+0.35	15.7	0.677	0.603						
	gemma-3-27b	+0.42	20.0	0.737	0.652						
	gpt-oss-20b	-1.21	11.8	0.656	0.595						
	Qwen3.5-9B	–	11.3	0.424	0.465						
	tgemma-4b	-2.85	3.8	0.601	0.554						
	Llama-3.1-8B	-2.05	4.3	0.513	0.500						
	gemma-3-4b	-3.35	5.4	0.526	0.510						

Table 11: Side-by-side performance breakdown for the **EN**→**TR** direction across subsets of TurkDoc-MT. ‘Hum’ denotes expert human negated MQM z-score, ‘xCOM’ stands for xCOMET-XL, and ‘TurC’ is TurCOMET. A dash indicates that no translation from that cell was drawn into the human audit sample.

Dataset	Model	Hum	BLEU	xCOM	TurC	Dataset	Model	Hum	BLEU	xCOM	TurC
FLORES	gemma-4-31B	+0.42	40.7	0.918	0.782	NTREX	gemma-4-31B	+0.42	34.5	0.705	0.649
	tgemma-27b	+0.36	39.0	0.912	0.777		tgemma-27b	+0.18	27.3	0.707	0.650
	tgemma-12b	+0.42	35.3	0.909	0.775		tgemma-12b	+0.42	31.4	0.677	0.629
	gpt-oss-120b	+0.42	37.1	0.895	0.763		gpt-oss-120b	+0.13	27.9	0.680	0.633
	gemma-3-12b	+0.42	38.6	0.905	0.771		gemma-3-12b	+0.42	32.0	0.669	0.623
	gemma-3-27b	+0.27	40.4	0.891	0.757		gemma-3-27b	+0.42	33.4	0.683	0.635
	gpt-oss-20b	+0.42	36.5	0.878	0.747		gpt-oss-20b	+0.33	27.0	0.656	0.617
	Qwen3.5-9B	+0.42	34.8	0.903	0.770		Qwen3.5-9B	+0.31	27.0	0.655	0.614
	tgemma-4b	+0.42	32.8	0.891	0.759		tgemma-4b	+0.22	27.4	0.647	0.614
	Llama-3.1-8B	+0.42	34.8	0.866	0.738		Llama-3.1-8B	+0.28	29.0	0.624	0.590
	gemma-3-4b	+0.12	33.0	0.861	0.733		gemma-3-4b	+0.23	26.3	0.611	0.585
TWiST	gemma-4-31B	+0.37	43.3	0.892	0.764	AYM	gemma-4-31B	+0.42	34.7	0.853	0.741
	tgemma-27b	+0.42	41.4	0.884	0.757		tgemma-27b	+0.42	30.5	0.851	0.742
	tgemma-12b	+0.30	36.8	0.879	0.753		tgemma-12b	+0.42	25.8	0.834	0.728
	gpt-oss-120b	+0.42	38.0	0.878	0.750		gpt-oss-120b	+0.42	27.8	0.844	0.734
	gemma-3-12b	+0.42	39.3	0.878	0.751		gemma-3-12b	+0.42	28.1	0.836	0.727
	gemma-3-27b	+0.35	41.1	0.870	0.743		gemma-3-27b	+0.42	30.0	0.825	0.716
	gpt-oss-20b	+0.11	37.0	0.867	0.744		gpt-oss-20b	+0.42	27.8	0.813	0.712
	Qwen3.5-9B	+0.42	37.6	0.883	0.756		Qwen3.5-9B	+0.33	29.8	0.832	0.727
	tgemma-4b	+0.42	33.4	0.857	0.735		tgemma-4b	+0.42	23.0	0.810	0.710
	Llama-3.1-8B	+0.42	35.5	0.849	0.726		Llama-3.1-8B	+0.15	28.9	0.789	0.691
	gemma-3-4b	+0.42	33.2	0.836	0.716		gemma-3-4b	+0.42	25.2	0.781	0.682
XQuAD	gemma-4-31B	+0.42	51.6	0.797	0.717	DocHPLT	gemma-4-31B	+0.42	43.2	0.770	0.688
	tgemma-27b	+0.42	47.3	0.792	0.715		tgemma-27b	+0.39	36.2	0.778	0.697
	tgemma-12b	+0.42	43.4	0.760	0.695		tgemma-12b	+0.37	32.6	0.765	0.687
	gpt-oss-120b	+0.42	44.4	0.772	0.696		gpt-oss-120b	+0.42	35.8	0.761	0.683
	gemma-3-12b	+0.42	46.3	0.760	0.687		gemma-3-12b	+0.37	38.8	0.758	0.674
	gemma-3-27b	+0.42	48.8	0.780	0.701		gemma-3-27b	+0.31	39.8	0.757	0.679
	gpt-oss-20b	+0.42	43.9	0.759	0.685		gpt-oss-20b	+0.02	34.1	0.746	0.672
	Qwen3.5-9B	+0.42	45.2	0.767	0.700		Qwen3.5-9B	+0.08	35.6	0.751	0.674
	tgemma-4b	+0.33	38.9	0.756	0.683		tgemma-4b	+0.22	33.7	0.753	0.676
	Llama-3.1-8B	+0.42	40.5	0.733	0.674		Llama-3.1-8B	+0.19	31.2	0.722	0.654
	gemma-3-4b	+0.42	40.4	0.724	0.661		gemma-3-4b	+0.27	34.1	0.729	0.656
TED2020	gemma-4-31B	+0.41	36.2	0.840	0.742						
	tgemma-27b	+0.19	19.2	0.780	0.704						
	tgemma-12b	-1.15	15.2	0.693	0.645						
	gpt-oss-120b	+0.42	24.7	0.799	0.713						
	gemma-3-12b	-0.12	30.7	0.758	0.684						
	gemma-3-27b	+0.42	33.6	0.832	0.734						
	gpt-oss-20b	+0.40	28.9	0.792	0.706						
	Qwen3.5-9B	+0.42	29.5	0.776	0.697						
	tgemma-4b	-2.31	8.6	0.742	0.670						
	Llama-3.1-8B	+0.42	22.8	0.764	0.685						
	gemma-3-4b	-1.84	18.0	0.456	0.501						

Table 12: Side-by-side performance breakdown for the **TR**→**EN** direction across subsets of TurkDoc-MT. 'Hum' denotes expert human negated MQM z-score, 'xCOM' stands for xCOMET-XL, and 'TurC' is TurCOMET.

Dataset	Direction	n	Metric	Pearson	Spearman	Kendall
FLORES	EN→TR	41	BLEU	+0.227	+0.146	+0.090
			xCOMET-XL	+0.378	+0.341	+0.247
			TurCOMET	+0.430	+0.373	+0.272
FLORES	TR→EN	33	BLEU	-0.018	-0.078	-0.064
			xCOMET-XL	+0.061	-0.079	-0.064
			TurCOMET	+0.064	-0.062	-0.050
TWiST	EN→TR	42	BLEU	+0.642	+0.694	+0.518
			xCOMET-XL	+0.579	+0.472	+0.361
			TurCOMET	+0.460	+0.447	+0.335
TWiST	TR→EN	39	BLEU	+0.312	+0.064	+0.049
			xCOMET-XL	+0.759	+0.273	+0.220
			TurCOMET	+0.518	+0.279	+0.226
XQuAD	EN→TR	33	BLEU	-0.408	-0.379	-0.267
			xCOMET-XL	-0.089	-0.164	-0.085
			TurCOMET	-0.033	-0.119	-0.063
XQuAD	TR→EN	36	BLEU	-0.102	-0.089	-0.074
			xCOMET-XL	-0.009	+0.008	+0.007
			TurCOMET	+0.060	+0.122	+0.101
NTREX	EN→TR	41	BLEU	+0.570	+0.594	+0.436
			xCOMET-XL	+0.466	+0.531	+0.380
			TurCOMET	+0.371	+0.372	+0.255
NTREX	TR→EN	43	BLEU	+0.311	+0.161	+0.133
			xCOMET-XL	+0.347	+0.221	+0.182
			TurCOMET	+0.282	+0.209	+0.168
AYM	EN→TR	36	BLEU	-0.079	+0.019	+0.020
			xCOMET-XL	+0.341	+0.373	+0.288
			TurCOMET	+0.201	+0.335	+0.257
AYM	TR→EN	38	BLEU	-0.202	-0.196	-0.163
			xCOMET-XL	-0.103	-0.173	-0.142
			TurCOMET	-0.123	-0.179	-0.146
DocHPLT	EN→TR	38	BLEU	+0.286	+0.243	+0.198
			xCOMET-XL	+0.613	+0.620	+0.479
			TurCOMET	+0.568	+0.559	+0.415
DocHPLT	TR→EN	46	BLEU	-0.563	-0.623	-0.489
			xCOMET-XL	-0.325	-0.407	-0.310
			TurCOMET	-0.223	-0.286	-0.212
TED2020	EN→TR	39	BLEU	+0.769	+0.684	+0.520
			xCOMET-XL	+0.542	+0.486	+0.361
			TurCOMET	+0.508	+0.555	+0.411
TED2020	TR→EN	45	BLEU	+0.595	+0.369	+0.284
			xCOMET-XL	+0.446	+0.488	+0.381
			TurCOMET	+0.714	+0.553	+0.439

Table 13: Segment-level correlations between each automatic metric and the z-normalised mean MQM penalty, broken down by dataset and translation direction. With only $n = 33\text{--}46$ rated translations per cell ($\text{CI} \approx \pm 0.32$ on a correlation), the per-domain differences between metrics are not statistically significant; bolding marks the numerically highest value only. See Section 6.7 for the pooled per-direction comparison and significance tests.

G Author Contributions

Atahan Uz designed and implemented the TurCOMET architecture, conducted model training and ablation studies, managed the LLM zero-shot inference pipelines, implemented the GPT-5.5 SQM automated filtering, executed all statistical and meta-evaluation analyses, and prepared the open-source release of the models and codebase.

Hüseyin Emir Akdağ led the curation and construction of the multi-domain TurkDoc-MT benchmark, implemented the sentence-level alignment filtering for the web crawl data, recruited the domain-expert annotators, managed the human quality audit process, and assisted in setting up the MQM annotation interface.

Susan Üsküdarlı supervised the research, providing strategic guidance on the methodological framework and evaluation design, contributing to the inter-annotator agreement calculations and tie-breaking mechanisms, and offering critical review and editing of the manuscript.