Professional Translators Versus Quality Estimation Models: Reliability and Agreement in English-Ukrainian Translation Evaluation

Dmytro Chaplynskyi
Ukrainian Catholic University
I. Krypiakevych Institute
of Ukrainian Studies, lang-uk
chaplynskyi.dmytro@ucu.edu.ua

Kyrylo Zakharov
UNHCR
kirillzakharov13@gmail.com

Lesia Ivashkevych
NTUU "Igor Sikorsky Kyiv
Polytechnic Institute"
lesia.ivashkevych@gmail.com

Abstract

We extend a prior study comparing automatic Quality Estimation (QE) models with crowdsourced student judgments for English–Ukrainian parallel corpus evaluation. Eight professional translators each rate 1,000 sentence pairs on a continuous 0–100 scale under one of two paradigms: holistic quality scoring or a two-stage fluency-plus-adequacy protocol, with a repeated task for test–retest reliability. Professionals using the holistic scale achieve significantly higher inter-rater reliability than both linguistics students and professionals using separate fluency and adequacy scales, contradicting the expectation that multidimensional evaluation improves agreement. Adequacy correlates strongly with holistic judgments while fluency emerges as a largely independent dimension. Experts also exhibit a significant leniency drift over the session, alongside increasing evaluation speed. We additionally evaluate three LLMs as translation quality judges (Gemini 3 Flash, GPT-5.4, Gemma 3 27B) and find that the two larger models modestly outperform dedicated QE models in correlation with expert scores ($r = 0.814$–0.821 vs. $r \leq 0.747$). When prompted for separate fluency and adequacy scores, the LLMs replicate the adequacy-dominance pattern, confirming that meaning preservation drives holistic quality perception across both human and machine judges.

1 Introduction

Evaluating the quality of machine translation (MT) output is essential for building and filtering parallel corpora, particularly for low- to mid-resource languages where training data quality directly affects downstream model performance. Automatic Quality Estimation (QE) models predict translation quality from the source–target pair alone, without requiring a human reference; this distinguishes them from reference-based metrics such as BLEU or chrF and makes them practical for corpus filtering at scale. Recent neural QE systems—COMET (Rei et al., 2020), COMETKiwi (Rei et al., 2022, 2023), xCOMET (Guerreiro et al., 2024), and MetricX (Juraska et al., 2023, 2024)—have become standard tools for this task, offering scalable sentence-level quality predictions.

However, the relationship between automatic QE scores and human quality perception is not straightforward. In a prior study (Chaplynskyi and Zakharov, 2025), we applied six QE models to score 55 million English-Ukrainian sentence pairs and conducted a human evaluation of a stratified sample of 9,775 pairs using linguistics students as annotators. The best ensemble model explained approximately 60% of the variance in averaged human ratings, with a non-linear relationship between automatic scores and human perception. Critically, inter-rater agreement among students was only moderate (ICC = 0.43–0.54), raising the question of whether limited agreement reflects genuine subjectivity in quality assessment or insufficient evaluator expertise.

The present paper addresses this question directly. We make four contributions:

1. **Professional evaluation at scale.** We recruit eight professional translators and collect over 8,000 individual evaluations on 1,000 sentence pairs, enabling direct comparison with student data from our prior study.

2. **Holistic versus multidimensional evaluation.** We employ two paradigms in parallel: a single holistic quality score and a two-stage fluency-plus-adequacy protocol (Graham et al., 2013; Lommel et al., 2014). Contrary to expectations, holistic scoring yields higher inter-rater agreement.

3. **Systematic reliability analysis.** We examine inter-rater and test–retest reliability, evaluator learning effects, leniency drift, and the role

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 97–107

of text complexity and evaluation time in explaining score variance.

4. **LLM-as-a-Judge comparison.** We evaluate three LLMs as translation quality judges and find that the two larger models outperform dedicated QE models in agreement with professional translators, while a smaller model does not benefit from structured prompting.

2 Related Work

Human evaluation of MT quality. The dominant paradigms for human MT evaluation include Direct Assessment (DA) on a continuous 0–100 scale (Graham et al., 2013, 2016), the Multidimensional Quality Metrics (MQM) framework based on error annotation (Lommel et al., 2014), and pairwise ranking. DA with continuous scales has been adopted by WMT shared tasks as the standard for collecting human judgments at scale (Freitag et al., 2021, 2022). The separation of fluency and adequacy as distinct evaluation dimensions has a long history in MT evaluation (Castilho et al., 2017; Görög, 2014), though recent work has debated whether holistic scoring produces comparable results with lower annotator burden.

Evaluator expertise. The effect of annotator expertise on MT evaluation reliability remains underexplored. Freitag et al. (2021) showed that expert evaluators produced higher agreement and different system rankings compared to crowdsourced judgments. Our prior study (Chaplynskyi and Zakharov, 2025) used linguistics students, achieving moderate inter-rater reliability (ICC = 0.43–0.54). The present study extends this by recruiting professional translators to test whether domain expertise improves reliability.

Quality estimation models. Neural QE models have advanced rapidly. The COMET family (Rei et al., 2020, 2022, 2023) combines COMET's architecture with OpenKiwi's predictor-estimator setup. xCOMET (Guerreiro et al., 2024) adds error span detection. The MetricX family (Juraska et al., 2023, 2024) uses a two-stage fine-tuning strategy on human-labeled data. These models achieve high agreement with human judgments on high-resource language pairs (Freitag et al., 2022), but their performance on low-resource pairs such as English-Ukrainian is less established.

LLM-as-a-Judge. Recent work has demonstrated that large language models can serve as effective MT evaluators. Kocmi and Federmann (2023) showed that GPT-based evaluation achieves state-of-the-art correlation with human judgments, and Zheng et al. (2023) established the LLM-as-a-Judge paradigm more broadly. We evaluate three LLMs as translation quality judges: Gemma 3 27B (Gemma Team, 2025), Gemini 3 Flash, and GPT-5.4, alongside traditional QE models.

3 Methodology

3.1 Data Sample

The evaluated texts were drawn from the OPUS Open Parallel Corpora (Tiedemann, 2016) and consist of 1,000 English-Ukrainian sentence pairs. Of these, 720 pairs overlap with the sample evaluated by linguistics students in Chaplynskyi and Zakharov (2025), enabling direct cross-study comparison. The remaining 280 pairs were randomly selected from the same corpus under the constraint that they were not corrupted and did not contain inappropriate or sensitive content.

3.2 Expert Recruitment

Nine professional translators were recruited, all with a minimum of three years of professional experience translating from English into Ukrainian. Candidates were identified through professional networks, with preference for translators whose reliability could be vouched for by colleagues. The translators' professional backgrounds span technical, legal, literary, media, marketing, and military translation domains, providing diversity in evaluation perspectives.

Translators were randomly assigned to two groups: four evaluated using a holistic quality scale (Group 1), and five evaluated using separate fluency and adequacy scales (Group 2). One translator in Group 2 completed only 39 evaluations and was excluded from the analysis, leaving four active evaluators per group and eight professional translators in total.

3.3 Evaluation Protocol

Evaluations were collected using Vulyk,[1] an open-source crowdsourcing platform, with two task plugins developed for this study: one for holistic transla-

[1] https://github.com/mrgambal/vulyk/

tion evaluation[2] and one for the two-stage fluency-and-adequacy protocol.[3] Twenty evaluation tasks were constructed, each containing 50 translation pairs sampled without replacement from the pool of 1,000 pairs. The order of pairs within each task was randomized. To assess test–retest reliability, the first task was repeated at the end of the evaluation sequence. Before beginning, all evaluators were shown five worked examples. Breaks were permitted between tasks but not during a 50-pair task. Timestamps were recorded for each evaluation from presentation to submission.

Holistic evaluation (Group 1). Evaluators were presented simultaneously with the English source text and its Ukrainian translation and instructed: "Rate the translation (0–100)." Ratings were provided using a continuous slider with a red-to-green color gradient and the following descriptors: 0–10 *Incorrect translation*, 11–29 *A few correct keywords, but the meaning is different*, 30–50 *Major mistakes in translation*, 51–69 *Understandable but contains typos or grammatical errors*, 70–90 *Preserves semantics closely*, 91–100 *Perfect translation*. This replicates the protocol used for students in the prior study (Chaplynskyi and Zakharov, 2025).

Fluency and adequacy evaluation (Group 2). Evaluators assessed translations in two successive stages. In the fluency stage, only the Ukrainian translation was displayed, and evaluators rated linguistic quality on a 0–100 scale (0–25 *Incomprehensible*, 25–50 *Disfluent*, 50–75 *Good*, 75–100 *Flawless*). After completing the fluency rating, the English source was revealed, and evaluators rated adequacy: "How much of the meaning expressed in the source text is also expressed in the target translation?" (0–25 *None*, 25–50 *Little*, 50–75 *Most*, 75–100 *All*). The sequential design prevents the source text from biasing fluency judgments.

3.4 Automatic Quality Estimation

Machine translation quality was assessed using nine automatic metrics. Six are dedicated QE models from two families: the COMET family (wmt22-cometkiwi-da, wmt23-cometkiwi-da-xl, wmt23-cometkiwi-da-xxl, and xCOMET-XXL) and the MetricX family (MetricX-23 and MetricX-24). We also include bicleaner-ai (Zaragoza-Bernabeu et al., 2022), a parallel-corpus cleaning classifier trained to flag noisy sentence pairs; cosine similarity of LaBSE sentence embeddings (Feng et al., 2022), a multilingual sentence encoder that supports cross-lingual semantic similarity; and Gemma 3 27B (Gemma Team, 2025) as an LLM-as-a-Judge baseline using a holistic scoring prompt matching the human evaluation rubric.

MetricX scores were rescaled from their native 0–25 inverted scale to 0–1 using $\text{score}_{\text{adj}} = 1 - \text{score}/25$. Gemma 3 scores were rescaled from 0–100 to 0–1.

3.5 Statistical Analysis

Expert scores were analyzed in raw, z-score-normalized, and percentile-rank-transformed forms. We report results primarily on z-score-normalized data, as this transformation yielded the most precise reliability estimates.

Inter-rater reliability. We summarised inter-rater agreement using the Intraclass Correlation Coefficient (ICC), which expresses the share of total score variance attributable to genuine differences between sentence pairs as opposed to disagreement among evaluators. We estimated ICC with a two-way random-effects, absolute-agreement, single-measures model (Shrout and Fleiss, 1979; Koo and Li, 2016) and interpreted values following Cicchetti (1994): below 0.40 = poor, 0.40–0.59 = fair, 0.60–0.74 = good, 0.75–1.00 = excellent.

Test–retest reliability. For individual evaluators, test–retest ICC was computed on the 50 pairs evaluated twice during the first and last task (Gisev et al., 2013). Systematic bias was assessed using paired t-tests.

Mixed-effects models. To test hypotheses about evaluation dynamics, we fitted linear mixed-effects models using lme4 (Bates et al., 2015), with random intercepts for sentence pairs and evaluators. Evaluations exceeding 120 seconds were excluded as likely reflecting breaks rather than sustained attention.

Text complexity. Readability indices (ARI, Coleman-Liau, FORCAST, nWS, RIX) and lexical diversity measures were computed on the English source texts using quanteda.

[2] https://github.com/lang-uk/vulyk-translations

[3] https://github.com/lang-uk/vulyk-fluency-adequacy

Expert	Type	N	Mean	Med.	SD
Exp. 1	Student	709	9.8	7.8	7.9
Exp. 2	Student	707	12.6	9.7	11.0
Exp. 3	Student	705	29.5	25.5	18.0
Exp. 4	Holistic	1017	15.4	12.1	13.4
Exp. 5	Holistic	911	43.5	37.4	26.7
Exp. 6	Holistic	933	19.9	16.0	16.0
Exp. 7	Holistic	889	41.5	33.6	27.2
Exp. 8	F&A	1009	28.0	24.8	16.6
Exp. 9	F&A	924	38.6	29.1	28.0
Exp. 10	F&A	993	34.4	30.8	17.9
Exp. 12	F&A	1011	20.3	16.8	14.3

Table 1: Evaluation duration in seconds per evaluator (evaluations $<$120 s). F&A = fluency and adequacy.

Expert	Scale	ICC	95% CI	Bias
Exp. 4	Holistic	0.21	(−0.07; 0.46)	−10.5
Exp. 5	Holistic	0.55	(0.32; 0.72)	16.7
Exp. 6	Holistic	0.58	(0.36; 0.74)	4.9
Exp. 7	Holistic	0.83	(0.72; 0.90)	12.2
Average	Holistic	0.85	(0.75; 0.91)	5.8
Exp. 8	Fluency	0.69	(0.51; 0.81)	5.8
Exp. 9	Fluency	0.78	(0.64; 0.87)	17.6
Exp. 10	Fluency	0.65	(0.46; 0.79)	18.1
Exp. 12	Fluency	0.76	(0.61; 0.86)	7.2
Average	Fluency	0.86	(0.77; 0.92)	12.2
Exp. 8	Adequacy	0.62	(0.42; 0.77)	9.6
Exp. 9	Adequacy	0.62	(0.42; 0.77)	12.6
Exp. 10	Adequacy	0.46	(0.21; 0.66)	16.6
Exp. 12	Adequacy	0.63	(0.43; 0.77)	14.7
Average	Adequacy	0.79	(0.66; 0.88)	13.4

Table 2: Test–retest reliability for individual experts (50 repeated pairs). Bias = mean score increase from first to last session. All paired t-tests significant ($p < 0.05$).

Group	ICC	95% CI	
Experts (holistic)	0.72	0.69	0.74
Adequacy	0.66	0.64	0.69
Fluency	0.59	0.56	0.62
Students	0.57	0.53	0.61
COMETKiwi+wmt22	0.86	0.84	0.87
xCOMET+MetricX-24	0.80	0.77	0.83
Machines (no bicleaner)	0.69	0.67	0.72
LLM judges	0.82	0.76	0.86
LLM fluency	0.68	0.57	0.75
LLM adequacy	0.80	0.72	0.85

Table 3: Group-level ICC (two-way, agreement, single measures) on z-score-normalized ratings.

4 Results

4.1 Descriptive Statistics

The dataset comprises 10,566 individual evaluations: 2,127 from three linguistics students (prior study), 4,200 from four professional translators using the holistic scale, and 4,239 from four professional translators using the fluency and adequacy scales. Table 1 summarizes the evaluation time by evaluator.

Median evaluation times range from 7.8 s (fastest student) to 37.4 s (slowest holistic expert). Professional translators in both conditions are generally slower than students, consistent with more deliberate evaluation.

4.2 Test–Retest Reliability

Individual test–retest ICC values, computed on the 50 pairs evaluated in both the first and last task, are reported in Table 2. For all experts, the time between the first and the last batch was no less than six days, except for Expert 12, who completed all tasks within two days. Previous methodological research suggests that test–retest intervals are typically chosen to balance recall bias and true change, most commonly ranging from a few days to approximately two weeks (Marx et al., 2003). Nevertheless, the re-test ICC for Expert 12 was not statistically different from the ICCs of the other experts.

Several patterns emerge. First, Expert 4 demonstrates poor test–retest reliability (ICC = 0.21, CI includes zero), suggesting inconsistent scoring behavior. Second, all experts show a positive bias—scores in the repeated task are higher than in the initial task—indicating a systematic leniency drift over the evaluation session. Third, averaged expert scores yield good to excellent reliability (ICC = 0.79–0.86), substantially exceeding individual reliability, confirming that aggregation across evaluators stabilizes judgments. Fourth, fluency test–retest reliability (average ICC = 0.86) is numerically higher than adequacy (0.79), though the difference is not statistically significant given overlapping confidence intervals.

4.3 Inter-Rater Reliability

We computed group-level ICC for each evaluator category using standard score ratings (z-score), which we determined to yield the most precise reliability estimates across normalization methods tested (raw, percentile, rank-based inverse normal transformation).

Figure 1 and Table 3 report the group-level ICC values for each evaluator category using z-score-normalized scores.

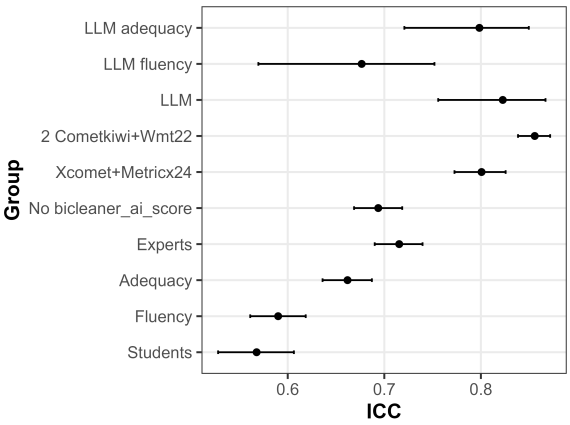

Figure 1: ICC with 95% confidence intervals for evaluator groups (z-score normalized).

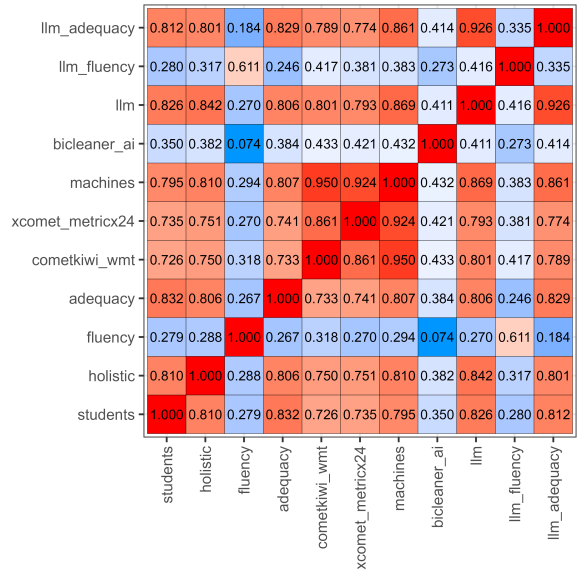

Figure 2: Correlation matrix between averaged group scores and machine model scores (z-score normalized).

Hypothesis 1: Professional translators show higher inter-rater reliability than students. Professional translators using the holistic scale achieved significantly higher inter-rater agreement (ICC = 0.72, 95% CI 0.69–0.74) than linguistics students (ICC = 0.57, 95% CI 0.53–0.61): the confidence intervals do not overlap. This confirms that evaluator expertise matters for translation quality assessment and that the moderate agreement observed in Chaplynskyi and Zakharov (2025) was at least partly attributable to evaluator inexperience rather than inherent task subjectivity.

Hypothesis 2: Multidimensional evaluation yields higher agreement. Contrary to our hypothesis, the holistic single-score evaluation (ICC = 0.72) yielded higher inter-rater agreement than either the adequacy (ICC = 0.66) or fluency (ICC = 0.59) scales. The adequacy ICC confidence interval (0.64–0.69) does not overlap with the holistic group (0.69–0.74), confirming that holistic evaluation produces significantly higher agreement. Fluency shows the lowest agreement among professional groups. This finding suggests that decomposing quality judgment into separate dimensions does not improve—and may slightly reduce—evaluator consistency, possibly because the cognitive task of isolating fluency from adequacy introduces additional judgment uncertainty.

Leave-one-out analysis. Leave-one-out ICC analysis confirmed that Expert 4 is an outlier: removing this evaluator substantially increases the holistic group's ICC. Expert 11 (already excluded due to insufficient evaluations) would similarly degrade the fluency-adequacy group if included. Among machine models, bicleaner-ai exhibits a distinct scoring pattern; excluding it increases the machine group ICC, consistent with its low correlation with other metrics observed in Chaplynskyi and Zakharov (2025).

4.4 Holistic Versus Multidimensional Scores

Hypothesis 3: Holistic scores correlate with adequacy more than with fluency. The correlation analysis confirms this hypothesis (Figure 2). Averaged holistic expert scores show a strong correlation with averaged adequacy scores ($r = 0.806$), while the correlation with fluency scores is notably weaker ($r = 0.288$). This indicates that when translators assign a single quality score, they weight meaning preservation (adequacy) more heavily than linguistic surface quality (fluency).

Fluency emerges as a largely independent dimension. Its highest correlation with any other group is only $r = 0.318$ (with COMETKiwi+wmt22), while even fluency–adequacy correlation is low ($r = 0.267$). This raises the question of whether fluency, as measured here, captures a quality dimension that is relevant to translation quality assessment in the context of parallel corpus filtering.

Student and expert holistic scores correlate strongly ($r = 0.810$), but expert scores show slightly stronger correlation with machine model scores ($r = 0.810$ vs. $r = 0.795$ for students), suggesting that professional judgment aligns more closely with what QE models capture.

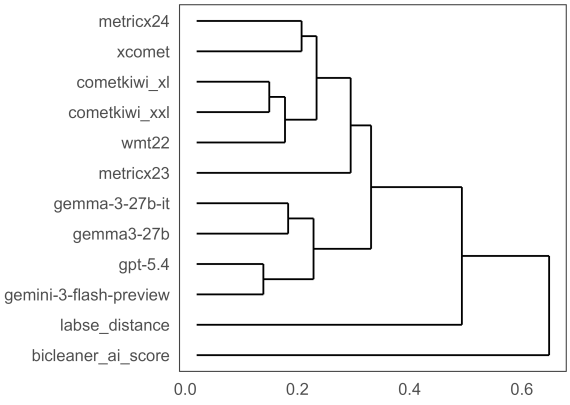

Figure 3: Hierarchical clustering of machine QE models and LLM judges based on Spearman correlation distance (z-score normalized scores).

4.5 Correlation with Automatic Metrics

The same non-linear (quadratic) relationship between machine and expert scores observed in Chaplynskyi and Zakharov (2025) persists in the professional evaluation data: QE models and human experts diverge most for translations of moderate quality, where automatic scores tend to be more optimistic than human judgments.

Among the machine models, the COMET family and xCOMET show the strongest correlations with expert holistic scores ($r = 0.750$ for COMETKiwi+wmt22, $r = 0.751$ for xCOMET+MetricX-24), while bicleaner-ai demonstrates weak correlations with all human and machine scores ($r = 0.382$ with experts). The hierarchical clustering of machine model scores (Figure 3) reveals two distinct clusters: one containing the COMETKiwi variants and wmt22-cometkiwi-da, and another containing xCOMET and MetricX-24. This clustering motivated our decision to compute group-level ICC separately for these model families (Table 3), as treating all machine models as a single group would conflate metrics that capture partially different aspects of translation quality.

An additional cluster composed of the LLM judges sits further from the traditional QE metrics, suggesting that the divergence reflects not only model differences but also the evaluation paradigm: LLM-based judgments depend on instruction-following and prompt design, introducing additional variability compared with regression-based QE metrics.

Model	N	r	**Mean**
LLM-as-a-Judge (rubric prompt, 0–100)			
Gemini 3 Flash	999	0.821	75.7
GPT-5.4	1000	0.814	68.1
Gemma 3 27B	1000	0.722	76.3
QE models			
Gemma 3 27B (QE)	710	0.747	—
COMETKiwi-XXL	710	0.740	—
xCOMET	710	0.735	—
Expert avg	1000	—	70.9

Table 4: Spearman correlation (r) of automatic scores with averaged expert holistic scores. Gemma 3 27B appears twice: as a QE baseline from the prior study (simple scoring, rescaled 0–1) and as an LLM-as-a-Judge with an explicit evaluation rubric (0–100 scale, temperature 0.3).

4.6 LLM-as-a-Judge

We evaluated three LLMs as translation quality judges on the same 1,000 sentence pairs: Gemma 3 27B (Gemma Team, 2025), Gemini 3 Flash, and GPT-5.4.[4] Each model scored all pairs on the 0–100 holistic scale. Gemini 3 Flash and GPT-5.4 additionally produced fluency and adequacy scores for all 1,000 pairs.

Table 4 reports the Spearman correlations between LLM scores and averaged expert holistic scores, with both variables converted to z-scores and means computed from raw scores. Gemini 3 Flash ($r = 0.821$) and GPT-5.4 ($r = 0.814$) slightly outperform all dedicated QE models in their baseline configuration from the prior study ($r = 0.747$). Notably, the Gemma 3 27B model yields consistent results across prompts, with $r = 0.747$ reported in (Chaplynskyi and Zakharov, 2025) and $r = 0.722$ in the current study. This suggests that a more detailed prompt with an explicit evaluation rubric (LLM-as-a-Judge) does not improve performance, indicating that a structured scoring prompt does not necessarily benefit smaller models.

GPT-5.4 is the best-calibrated model, with a mean score (68.1) closest to the expert average (70.9) and a score distribution that closely matches the expert distribution. Gemini and Gemma both exhibit the leniency bias observed across all LLM evaluators.

Fluency and adequacy raw scores from all three LLMs (1,000 pairs each) replicate the key patterns observed in human evaluation observed with

[4] All models were prompted with the same holistic scoring rubric used for human evaluators, with temperature set to 0.3.

Spearman correlations. LLM adequacy correlates strongly with expert holistic scores (Gemini: $r = 0.772$; GPT-5.4: $r = 0.759$; Gemma: $r = 0.704$) and expert adequacy (Gemini: $r = 0.809$; GPT-5.4: $r = 0.799$; Gemma: $r = 0.694$), while LLM fluency correlates weakly with expert holistic scores ($r = 0.227$–0.326) and shows low correlation with expert adequacy ($r = 0.174$–0.250). The internal fluency–adequacy correlations for Gemini ($r = 0.263$) and GPT-5.4 ($r = 0.288$) are comparable to human experts ($r = 0.267$), confirming that the models separate these dimensions similarly to professionals. Gemma shows weaker separation ($r = 0.447$). GPT-5.4 is also the best-calibrated model on the adequacy scale (mean 75.2 vs. expert 73.3).

Pairwise agreement among LLMs holistic scores is high (Gemini–GPT: $r = 0.861$; Gemma–GPT: $r = 0.798$; Gemini–Gemma: $r = 0.749$), indicating substantial convergence despite different architectures and training data.

4.7 Text Complexity and Score Variance

Hypothesis 4: Text complexity is positively correlated with score variance. We find no evidence for this hypothesis. Correlations between text complexity measures (readability indices, lexical diversity, sentence length) and the standard deviation of expert scores across evaluators are weak and inconsistent across evaluator groups. The only notable observation is that longer source texts are associated with lower fluency ratings, possibly because longer sentences provide more opportunities for grammatical or stylistic issues.

Hypothesis 7: Text complexity affects human scores more than machine scores. There is no strong evidence for differential impact. Text complexity measures show similarly weak relationships with score variance for both human and machine evaluators.

4.8 Evaluation Dynamics

Hypothesis 5: Experts assign higher scores over time. A linear mixed-effects model predicting the holistic score from evaluation order (with random intercepts for sentence pair and evaluator; $N = 5{,}871$ evaluations from 7 experts) reveals a small but significant positive effect of order on scores ($\beta = 0.006$, $SE = 0.002$, $t = 3.94$, $p < 0.001$). Over the course of approximately 1,000 evaluations, this corresponds to a cumulative increase of roughly 6 points on the 0–100 scale. This leniency drift is consistent with the positive biases observed in the test–retest analysis (Table 2) and suggests that experts become more lenient as they progress through the evaluation.

Hypothesis 6: Evaluation duration decreases over time. A mixed-effects model predicting evaluation duration from order ($N = 9{,}846$ evaluations from 12 evaluators) confirms a significant learning effect ($\beta = -0.012$, $SE = 0.001$, $t = -12.79$, $p < 0.001$). Experts become faster at the evaluation task as they gain experience, with the duration reduction amounting to approximately 12 seconds over 1,000 evaluations. This pattern is consistent across both evaluation conditions and indicates genuine task learning—experts develop more efficient evaluation strategies over time.

The combination of increasing speed and increasing leniency suggests a form of evaluator fatigue or habituation: as the task becomes more routine, experts spend less time on each pair and default to higher scores, potentially reflecting reduced attention to translation errors.

5 Discussion

The present study extends Chaplynskyi and Zakharov (2025) by introducing professional translators as evaluators and comparing holistic and multidimensional evaluation paradigms for English-Ukrainian translation quality assessment. Our findings offer several insights with implications for both MT evaluation methodology and practical corpus filtering.

Expertise matters, but method matters more. Professional translators using a holistic scale achieve substantially higher inter-rater reliability than linguistics students evaluating the same translations with the same interface and instructions. This confirms that the moderate agreement reported in our prior work was partly an artifact of evaluator inexperience. However, the improvement is specific to the holistic condition: professionals using separate fluency and adequacy scales do not clearly outperform students. This suggests that evaluation reliability depends on the interaction between evaluator expertise and task design, not on expertise alone.

The fluency-adequacy decomposition does not help. Our most surprising finding is that the multidimensional evaluation paradigm, despite its the-

oretical appeal and widespread use in MT evaluation, does not improve inter-rater agreement over simple holistic scoring. One possible explanation is that fluency and adequacy are not orthogonal for web-crawled parallel data of the type in our sample: most low-quality translations fail on both dimensions simultaneously (garbled text is neither fluent nor adequate), while most high-quality translations succeed on both. The decomposition may add value primarily for translations where fluency and adequacy diverge—a relatively rare case in naturally occurring parallel corpora. The fluency evaluation also carries the additional cognitive burden of evaluating a text without its source, which may introduce uncertainty.

Adequacy dominates quality perception. When translators assign a single quality score, their judgment aligns strongly with adequacy (meaning preservation) and only weakly with fluency (linguistic surface quality). This finding has practical implications for corpus filtering: if the goal is to predict human quality judgments, adequacy-oriented metrics may be more informative than fluency-oriented ones. It also suggests that the fluency dimension, at least as measured with the scale descriptors used here, captures something that human evaluators do not strongly weight when making holistic quality judgments about parallel corpus data.

Evaluator drift is real and should be monitored. The significant leniency drift we observe—experts assigning progressively higher scores over the evaluation session—is a practical concern for any large-scale human evaluation campaign. The concurrent decrease in evaluation time suggests that the drift reflects reduced engagement rather than genuine recalibration. Future evaluation protocols should consider randomizing the presentation order more aggressively, inserting calibration anchors throughout the session, or normalizing scores within evaluation blocks to mitigate this effect.

LLMs as viable replacements for QE models. The two largest LLM judges—Gemini 3 Flash ($r = 0.821$) and GPT-5.4 ($r = 0.814$)—outperform all dedicated QE models in correlation with expert holistic scores, both individually and as an averaged ensemble ($r \leq 0.810$). Gemma 3 27B achieves slightly lower performance as an LLM-as-a-Judge ($r = 0.722$) and in its QE configuration ($r = 0.747$), suggesting that structured prompting does not benefit smaller models. GPT-5.4 achieves the closest calibration to human experts (mean 68.1 vs. 70.9), while Gemini achieves the highest correlation. The LLMs also replicate the adequacy-dominance pattern: their QE scores correlate strongly with expert adequacy ($r = 0.69$–0.79) but weakly with expert fluency ($r = 0.22$–0.26), confirming that holistic quality perception—whether by humans or LLMs—is primarily driven by meaning preservation. When prompted for separate fluency and adequacy scores, Gemini and GPT-5.4 achieve fluency–adequacy separations ($r = 0.26$–0.29) comparable to human experts ($r = 0.26$), indicating that LLMs can meaningfully decompose translation quality into independent dimensions.

Implications for corpus filtering. The non-linear relationship between QE model scores and expert judgments persists when moving from student to professional evaluators, confirming that this is a genuine feature of the QE-human relationship rather than an artifact of student evaluation quality. For practical corpus filtering, this reinforces the recommendation from Chaplynskyi and Zakharov (2025) to use non-linear ensemble models rather than raw QE scores when estimating human quality perception.

6 Conclusion

We presented a systematic comparison of professional human evaluation and automatic quality estimation for English-Ukrainian machine translation, testing seven hypotheses. Our key findings are:

- Professional translators using holistic scoring achieve significantly higher inter-rater reliability than linguistics students, confirming that evaluator expertise improves the quality of human reference data for MT evaluation.
- Contrary to expectations, holistic evaluation outperforms the fluency-adequacy decomposition in terms of inter-rater agreement, suggesting that simpler evaluation protocols may be preferable for corpus-level quality assessment.
- Adequacy strongly predicts holistic quality judgments, while fluency is a largely independent dimension—indicating that meaning preservation is the dominant factor in how translators perceive overall translation quality.

- Experts exhibit a significant leniency drift (higher scores over time) coupled with faster evaluation, pointing to habituation effects that should be accounted for in evaluation design.

- LLM-as-a-Judge evaluation with Gemini 3 Flash ($r = 0.821$) and GPT-5.4 ($r = 0.814$) outperforms all dedicated QE models in correlation with expert scores. All three LLMs replicate the adequacy-dominance and fluency-independence patterns observed in human evaluation.

For future work, we plan to: (1) evaluate the downstream impact of corpus filtering on NMT model performance; (2) investigate whether LLM-as-a-Judge models can fully replace human evaluation for corpus quality assessment at scale; and (3) apply the framework to other language pairs. We release the full sample of 1,000 evaluated English–Ukrainian sentence pairs together with all individual expert and LLM ratings[5] and the analysis code[6] to support reproducibility and external validation.

Limitations

- The study focuses on a single language pair (English-Ukrainian), and results may not generalize to other pairs with different morphological or resource characteristics.

- The sample of 1,000 sentence pairs, while sufficient for statistical analysis, represents a small fraction of the 55 million pairs in the full corpus.

- Professional translators may still exhibit domain-specific biases (e.g., technical vs. literary).

- One evaluator completed insufficient evaluations and was excluded; another (Expert 4) showed poor test–retest reliability (ICC = 0.21), suggesting that professional status alone does not guarantee evaluation quality.

- The fluency-adequacy evaluation inherently takes longer per pair (two ratings), which may introduce differential fatigue effects not present in the holistic condition.

- We acknowledge that the sequential fluency-then-adequacy design may introduce an anchoring effect, potentially influencing the independence of the two judgments. However, the observed low correlation suggests that any such bias did not artificially inflate agreement and is unlikely to have driven the main results.

- The evaluated sample is drawn from the publicly available OPUS corpus, which predates the training cutoffs of the LLMs we use as judges. We cannot rule out that some sentence pairs were seen during LLM pre-training, and a degree of memorization-driven inflation of LLM-as-a-Judge scores is therefore possible. The modest gap between the best LLM judge and the strongest dedicated QE model ($\Delta r \leq 0.07$ in Spearman correlation) makes contamination unlikely to be the sole driver of the observed advantage, but the effect cannot be quantified from the data available here.

Ethical Considerations

This study involves human evaluation of translation quality by professional translators who were compensated for their work. All evaluators participated voluntarily and were informed about the purpose of the study.

Parts of the codebase (data processing, analysis scripts, and the crowdsourcing platform plugins) were developed with the assistance of Claude Code (Anthropic), an AI-based coding tool. Claude Code was also used as a writing aid during the preparation of this manuscript. All AI-generated content was reviewed and edited by the authors. The LLM-as-a-Judge evaluation prompts are provided in Appendix A for reproducibility.

Acknowledgments

We thank the professional translators who contributed their time and expertise to this evaluation: Anton Shpigunov, Vladislav Demyanov, Kristina Zayka, Anatolii Zhylavyi, Oleksandra Vankevych, Faina Zholobak, Olena Pantsyr, and Olga Prokopchuk.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

[5] https://huggingface.co/datasets/lang-uk/qe-vs-human

[6] https://github.com/Amice13/translation-quality

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? In *The Prague Bulletin of Mathematical Linguistics*, volume 108, pages 109–120.

Dmytro Chaplynskyi and Kyrylo Zakharov. 2025. A framework for large-scale parallel corpus evaluation: Ensemble quality estimation models versus human assessment. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 73–85, Vienna, Austria (online). Association for Computational Linguistics.

Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, 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.

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*, pages 46–68. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Natasa Gisev, J Simon Bell, and Timothy F Chen. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338.

Attila Görög. 2014. Quantifying and benchmarking quality: The TAUS dynamic quality framework. *Tradumatica: Tecnologies de la Traducció*, (12):443–454.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. 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.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

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.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Arle Richard Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumatica: Tecnologies de la Traducció*, (12):455–463.

Robert G Marx, Alia Menezes, Lois Horovitz, Edward C Jones, and Russell F Warren. 2003. A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of clinical epidemiology*, 56(8):730–735.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702. 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. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645. Association for Computational Linguistics.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Jörg Tiedemann. 2016. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, page 384.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 824–831. European Language Resources Association.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

A LLM-as-a-Judge Prompts

Holistic QE prompt. All three LLMs received the following system prompt for holistic quality estimation:

> You are a professional English-to-Ukrainian translation evaluator. Rate the translation quality on a 0–100 scale using these descriptors: 0–10 Incorrect translation; 11–29 A few correct keywords, but the meaning is different; 30–50 Major mistakes in translation; 51–69 Understandable but contains typos or grammatical errors; 70–90 Preserves semantics closely; 91–100 Perfect translation. Return only a JSON object with a single "score" field.

Fluency and adequacy prompt. For the two-stage evaluation, the following prompt was used:

> You are a professional English-to-Ukrainian translation evaluator. First, evaluate the fluency of the Ukrainian text (0–100): 0–25 Incomprehensible; 25–50 Disfluent; 50–75 Good; 75–100 Flawless. Then, evaluate how much of the meaning from the English source is preserved in the Ukrainian translation (0–100): 0–25 None; 25–50 Little; 50–75 Most; 75–100 All. Return only a JSON object with "fluency" and "adequacy" fields.

All models were called with temperature 0.3. The English source text and Ukrainian translation were provided as user input.