

Looking under the Wrong Lamppost: On the Limitations of Automated Translation Quality Estimation

Serge Gladkoff
Logrus Global LLC

Angelika Vaasa
European Parliament,
DG for Translation and Clear Language

Sue Ellen Wright
Kent State University

Ingemar Strandvik
MQM Council

Lifeng Han
LUMC, Leiden & Leiden University

Abstract

Automation of Translation Quality Estimation (QE) has emerged as a widely discussed approach to managing translation quality at scale, and a growing number of tools and technologies have been released in pursuit of this goal. However, the proliferation of new QE systems has not always been accompanied by robust, transparent, and reproducible research and testing. This gap deserves critical scrutiny. This paper examines some fundamental limitations of the QE technology from both theoretical and empirical perspectives, arguing that current QE systems are structurally ill-equipped to serve as reliable standalone tools in real-world translation workflows. The reviewed evidence suggests that QE suffers from a range of interrelated and largely unresolved limitations. Most fundamentally, the evaluation of the quality of translation at the level of isolated segments is problematic because it tends to miss out on cohesion, coherence, and stylistic and rhetorical text features. In addition, empirical research documents several other limitations and flaws, including failure to generalize, systematic biases, overfitting and distribution collapse, performance gaps, error annotation challenges, and data scarcity.

Consequently, segment-level QE scores should not be used as a standalone basis for routing, release, or review bypass in production; we argue that future automation work should target automated MQM-based evaluation calibrated against human-based MQM. This paper does not advocate a return to exclusively human review.

1 Background

The translation industry has long grappled with the misconception that translation is a task that anyone with basic language skills can perform. As a discipline within language science, translation studies have always faced the fact that language

belongs to everyone because people use it daily for communication.

When a message is simple and clear, translation can indeed be a simple task. However, it can also require a high degree of sophistication. There is a long-standing body of work in translation studies that shows that certain linguistic and cultural phenomena are inherently resistant to direct translation (Catford, 1965; Jakobson, 1959).

Between these two extremes lies the full spectrum of content types and communicative modes that weave humanity into one interconnected world. Both realities coexist: some translation tasks are easy, while others are virtually unsolvable. This is precisely why relying solely on mechanical markers of quality, such as post-editing effort, fails to capture the full nuance of what translators do. As Yang et al. (2023) argue, word-level QE should be anchored in human semantic judgment rather than in the mere physical act of editing.

AI is helpful in many routine translation scenarios, even in crisis situations (Staiano et al., 2025), democratizing access to interlingual communication for a wide range of users. But AI also comes with a risk of capability misattribution: users perceive it as having capabilities that it does not possess. The shortcomings of machine translation (MT) and AI-assisted translation are not always self-evident. They are often detectable only by trained professionals who can identify and mitigate the significant and costly risks that arise from deploying unmanaged, unattended MT and AI in real-world settings. As Mehandru et al. (2023) demonstrate in a medical context, even when aided by QE, non-professional users of AI, such as physicians, may over-rely on technology and miss critical errors that can lead to clinical harm.

At the same time, risk-averse real-world process management, as mandated by ISO 9000, presumes that production workflows have been tuned to avoid

even relatively rare risk events.¹ Applying the same logic to translation production workflows means that MT and AI must be incorporated sensibly to achieve manufacturing-grade reliability, risk guarantees, and robustness. A QE approach that focuses exclusively on pre-production possibilities rather than post-production realities is difficult to reconcile with this requirement, as reflected by the findings of [Scarton et al. \(2019\)](#), and [Liu et al. \(2025\)](#).

With the advent of MT and AI translation, the industry has gained powerful new tools, but these tools are far from perfect, even for language pairs where English is one of the languages, let alone for other language pairs ([Han et al., 2024](#)). Researchers and practitioners increasingly recognize that the language problem is far from being solved ([Liang and Han, 2026](#)). Therefore, the industry needs to adopt a more rigorous and critical approach, one that prioritizes factual, reliable, and reproducible research.

Contributions. This paper makes three contributions. Conceptually, it identifies why sentence-level QE is limited when translation quality depends on extra-segmental context, formalized as $Q(s, t, c)$. Empirically, it analyzes 104,762 MQM-annotated segments across three quality regimes, separating statistical ranking performance from the stronger requirement of reliable triage. Practically, it argues that any decision use of such scores must be established by local validation rather than assumed.

2 Fundamental Flaw of the Segment-Based Approach

As research approaches the final, and arguably most difficult, frontier of language automation, it is important not to be drawn in by its promise alone, but to focus on where genuine solutions are most likely to emerge. Current research efforts may be disproportionately concentrated on areas that are computationally convenient rather than theoretically appropriate, or even pragmatically accurate.

Broad claims are made about the promise of QE systems. However, in professional, operational environments, the path from theoretical promise to real-world deployment requires rigorous, transparent, and reproducible experimental evidence to establish both safety and efficacy. In the absence of such evidence, or where evidence points to the

contrary, these claims should be treated with skepticism.

This situation evokes the classic streetlight effect—an observational fallacy where a person searches for lost keys under a lamppost simply because the light is better there, rather than in the dark alley where the keys were actually dropped. In automated evaluation, focusing exclusively on isolated, segment-level metrics represents searching where computation is easy and convenient, rather than venturing into the darker, more complex domain of extra-segmental context, linguistic multidimensionality, discourse structure, human perception, and communicative intent where true translation quality actually resides.

The notion of fully automated QE appears to conflict with fundamental properties of language, a fact that may be intuitive to professional translators or scholars of translation studies, but one that can also be demonstrated through simple, measurable analysis. Statistical and mathematical considerations suggest that assessing the translation quality of a single isolated sentence, without its contextual environment, is problematic ([Gladkoff et al., 2022](#)). This principle is also reflected in the Multi-Range Theory of Translation Quality Measurement advanced by the MQM Council ([Lommel et al., 2024](#)).

Constructing metrics to reliably represent translation quality at the level of isolated sentences is problematic because translation correctness depends strongly on extra-segmental context. A translation of a sentence can be simultaneously correct and incorrect depending on context, terminology settings, quality requirements, target audience, author intent, or source meaning that cannot be recovered from the sentence alone. In other words, extra-segmental factors are determinative. Even spelling and grammatical errors cannot always be reliably identified from a single segment in isolation.

Serious translation errors are distributed across the full range of segment-level QE scores ([Gladkoff, 2025](#)), as shown in Figure 2 (Appendix A). Consequently, those scores cannot serve in the production environment as a reliable basis for routing, auto-release or review bypass, and cannot be assumed to provide the stable, monotonic ordering of risk that prioritization of human review effort would require. At best, such scores may be retained for retrospective research or descriptive analytics, but should not serve as decision variables in pro-

¹<https://www.iso.org/standards/popular/iso-9000-family> specifically ISO 9001:2015

duction.

Proponents of segment-based evaluation sometimes argue that many contextual aspects can be derived from a training corpus. We argue that this is a misconception and one with a longer history in translation technology. Before the advent of CAT tools, translators translated thoughts, not sentences. When translation memories arrived, many translators objected to the idea of dissecting a text into constituent parts and then claiming a productivity gain by reassembling new content from old fragments. Their concern was that context was being lost and that even 100% translation-memory matches might require different translations in a new context (Dragsted, 2004; Ranasinghe et al., 2020). If we consider what translators actually do when working in CAT tools, we see that they are not merely verifying translation-memory matches. They are reconstructing the meaning of source content using fragments of previous translations stored segment by segment and stripped of their original contextual reference.

From a formal perspective, translation quality can be viewed as a function not only of the source and target segments, but also of contextual variables, such as discourse structure, terminology constraints, and communicative intent. Because these variables are not recoverable from isolated segments, any segment-level QE necessarily operates under incomplete information, which limits its validity as a standalone measure.

Translation quality could be conceptualized as a function $Q(s, t, c)$, where s denotes the source segment, t the target translation, and c the set of contextual parameters relevant to specifications-based quality assessment. Appendix B gives examples involving grammatical gender and domain-dependent terminology.

3 Limitations of Current QE Systems

3.1 Generalization and Distribution Shift

Overfitting and Distribution Collapse. When a QE model is trained on a corpus of past translations, it does not learn how to translate from this corpus. Rather, it learns the patterns of translations deemed correct within the available data, inevitably diluted by translation errors stored in the corpus itself.

Even when a QE model is trained on an exceptionally clean translation corpus, and even when training yields apparently strong results on in-domain test data, applying the model outside the

distribution of its training corpus can lead to substantial degradation in predictive reliability. A recent benchmark for English–Hebrew QE directly confirms the instability of QE models when applied outside their training distributions. Rosenbaum et al. (2026) introduced an MTQE.en-he benchmark and documented fine-tuning experiments using widely adopted models such as TransQuest and COMETKiwi. Their results exhibited sensitivity to overfitting and distribution collapse. This finding is particularly significant because it arises even with clean annotated data and pre-trained models backed by substantial computational resources.

This evidence suggests that even state-of-the-art QE systems cannot reliably generalize beyond their specific training contexts, validating long-standing concerns about QE instability in real-world scenarios characterized by linguistic and use-case diversity. Morphological complexity, a hallmark of many languages, appears to be a particularly challenging dimension for current QE models to navigate, suggesting that existing training procedures have limited ability to learn language-agnostic quality indicators (Chahuneau et al., 2013; Siani et al., 2026).

The limitation appears to hold even when the training corpus is very large. Gladkoff (2025) reports on a large-scale, real-world case study by Bane covering technical content in 9 languages, using an industry standard pre-trained QE model. The study found that segments containing major and critical errors received scores spanning the entire range of the model’s output—including the highest scores. In more than one language, critical errors appeared even among the top-scoring 5 sentence pairs out of a set of 2,000. When serious errors are distributed across the full score range in this way, the QE score becomes difficult to use in practice, as no threshold can reliably separate translations that are fit for use as-is from those that require human review (see Figure 2).

These limitations help explain why the field of translation QE continues to encounter significant challenges with respect to the robustness and stability of automatic evaluation systems, particularly at the segment and span levels. The findings of the WMT25 Shared Task also highlighted persistent performance gaps consistent with these larger concerns (Lavie et al., 2025). The same kind of limitations at the model level are reflected in the performance of downstream evaluation, as discussed below.

Domain Variation and Generalization Failures. One of the most critical vulnerabilities of QE systems is their inability to generalize across domains and data sources. Large cross-domain variations in auto-rater behavior are supported by multiple studies that document systematic domain generalization failures. [Kocyigit et al. \(2022\)](#) demonstrate this failure in parallel corpus mining settings, where models trained on specific datasets could not reliably assess quality in broader real-world contexts. [Sharami et al. \(2023\)](#) show that even domain-tailored models struggle to maintain robustness when confronted with text characteristics or language pairs outside their adaptation scope.

Synthetic Data Generation Issues. The reliance on synthetic data – machine-generated translations with pseudo-labels—for training QE systems introduces additional stability concerns. [Geng et al. \(2025\)](#) show that pseudo-labels frequently fail to align with human preferences. The distribution shift between synthetic pseudo-translations and real-world translations creates systematic mismatches that undermine model generalization.

The reliance on massive parallel corpora also creates what [Kocyigit et al. \(2022\)](#) describe as a structural feedback loop in which only organizations with large existing data resources can build the tools necessary to find more data. Current QE models therefore risk becoming over-optimized for the narrow distributions represented in their training sets.

Because high-quality human-annotated translation quality data are scarce and expensive, synthetic data is difficult to avoid in practice. However, this reliance introduces exactly the type of distribution shift that makes robust deployment difficult.

3.2 Annotation and Human–Machine Misalignment

Human–Machine Misalignment. A fundamental issue undermining QE stability is the misalignment between what human annotators label as errors and what QE systems learn to predict. [Yang et al. \(2023\)](#) find that traditional post-editing effort labels often conflict with human judgment about whether a word is actually well or poorly translated. This suggests that QE systems trained on conventional datasets may be optimizing for proxy metrics that do not closely correspond to true translation quality.

The problem is compounded by a misalignment between model targets and human workflow needs.

As [Sarti et al. \(2025a\)](#) demonstrate, traditional QE labels often conflict with post-editors’ actual preferences. Thus, even when QE models achieve high correlation scores against standard benchmarks, their predictions offer limited practical utility because they fail to capture what human editors genuinely need to correct.

This gap between human judgment and automatic systems is particularly pronounced at the word and span levels. [Sarti et al. \(2025b\)](#) show that relying on single-annotator evaluations masks underlying instability; when multiple sets of human labels are introduced, supervised methods struggle with severe label uncertainty. In classical terms, this highlights a fundamental issue of inter-rater reliability in fine-grained QE annotation ([Scarton et al., 2019](#)). [Mehandru et al. \(2023\)](#) also show in a high-stakes medical setting that QE can help identify some translation issues, but it still fails to reliably detect critical errors compared to more manual methods, such as back-translation.

Research on fine-grained error prediction confirms that identifying exact error boundaries and assessing severity remains highly challenging.

Taken together, these findings suggest that the usability of QE in professional post-editing workflows remains substantially limited. Domain, language pair, and editor expertise all emerge as critical factors in determining QE effectiveness, indicating that current QE systems are highly context-dependent, brittle, and unreliable.

3.3 Biases and Error Detection Limitations

Performance Gaps and LLM-based Auto-Rater Limitations. Recent work has shown that LLM-based auto-raters (sometimes called “LLM-as-a-judge”, although neither term defines the exact method precisely) typically underperform on fine-grained quality assessment tasks. The difficulty of identifying fine-grained error spans and accurately classifying their severity represents a fundamental limitation of current QE approaches. [Jung et al. \(2023\)](#) show that even when the task is decomposed into sequential stages –sentence-level detection followed by word-level span extraction—the individual components continue to struggle with accuracy. Studies on Chinese homophone translation also indicate that while larger LLMs exhibit greater stability and robustness to perturbations, they still show substantial vulnerabilities when evaluating nuanced semantic content ([Qian et al., 2025](#)).

A major obstacle to the adoption of QE in pro-

duction is its opaque nature. As [Dinh and Niehues \(2023\)](#) argue, most current QE systems fail to provide explainable feedback, making it difficult for human editors to trust or verify their outputs. Their perturbation-based QE framework points toward a possible shift toward unsupervised, explainable, and transparent methods that do not rely as heavily on biased training corpora, although it also illustrates how attempts to solve one set of QE problems can introduce additional complexity.

Error Detection Limitations. Auto-raters continue to struggle with precise error detection, span annotation, and severity classification, falling short of the precision required by professional MQM standards ([Jung et al., 2023](#)).

Extensive analysis of recent WMT shared tasks indicates that, while sentence-level QE has improved considerably, finer-grained tasks such as word-level QE and error span detection remain a challenge. [Qian et al. \(2025\)](#) show that larger LLMs are somewhat more stable, but still vulnerable when evaluating translations involving *nuanced semantic content*. Likewise, although xCOMET represents a substantial step toward transparency, practical implementation of such granular schemes still falls short of the MQM gold standard for reliably distinguishing error types and severities ([Guerreiro et al., 2023](#)).

Systematic Biases. Recent research has exposed systematic biases embedded in QE metrics themselves. [Zhang et al. \(2025\)](#) show that QE systems consistently assign lower quality scores to translations as they become longer, even when those translations are entirely accurate. This length bias arises because training data is skewed: models learn to associate length with error probability and therefore penalize paraphrastic translations, regardless of whether they are perfectly fit for their purpose.

Another documented bias concerns gender. [Zaranis et al. \(2025\)](#) show that, when gender is ambiguous in the source text, current QE models systematically favor masculine inflections with higher quality scores, regardless of actual translation accuracy. These findings suggest that QE metrics may conflate translation quality with superficial patterns in skewed training data.

Cross-System Heterogeneity and Inconsistency. Applying QE metrics to outputs from heterogeneous MT systems reveals further inconsistencies. The same QE model can produce markedly different results depending on the MT system being evaluated, suggesting that QE models may learn

system-specific regularities rather than generalizable quality indicators.

[Dinh and Niehues \(2023\)](#) partly addresses this issue through an explainable perturbation-based approach designed for black-box MT. Their work is valuable, but it also illustrates a broader challenge: QE often introduces new layers of complexity in the course of trying to solve older ones.

4 Empirical Validation with a Large Dataset

The empirical test from large industrial data reported in this section addresses the operational question that follows from the preceding discussion. That question is not whether a segment-level QE system can produce a score that differs from random, but whether such a score can be treated as a calibrated quality score, as a portable threshold, or at least as a reliable ranking signal for the prioritization of segments for review.

Data and the gold standard. The experiment used three MQM-annotated data sources representing different quality regimes. One data set was from the German Research Center for Artificial Intelligence (DFKI). The DFKI material consisted of legacy rule-based machine translation output with errors in nearly every sentence. The second dataset consisted of MT-generated, unedited translations from the European Parliament (EP). The third consisted of final human-edited translations from the European Commission (EC), representing a low-error production-quality regime. These three sources are important because they span the practical range in which QE is normally claimed to be useful: very poor MT output, an unedited AI/MT output from an engine considered good, and high-quality human-edited final translations.

Across the three primary corpora, the analysis covered **104,762 scored segments**: 2,543 DFKI segments, 56,715 EP segments, and 45,504 EC segments. All segments had undergone human MQM annotation, which was used as the gold standard. The main binary gold label was whether a segment contained any MQM error. A secondary major-plus AUC was also recorded where available, but the thresholded confusion matrices were generated for the any-error condition. This distinction matters because major-error counts were often sparse in the EC slices, making major-plus-threshold conclusions unstable.

QE model and analysis instrument. The segments were first scored using the TAUS EPIC QE model. The resulting scores and MQM annotations were analyzed using Revix, a statistical analysis platform designed for the evaluation of translation-quality workflows. Revix stores annotated translation datasets and provides statistical analyses of the relationship between QE outputs and human quality annotations, including ranking performance, calibration, threshold behavior, and workflow-level utility metrics. In this study, Revix was used solely as an analysis environment; the object of evaluation was the QE scores rather than the Revix platform itself.

TAUS EPIC AIQE was the first metric scored on this material; scores from five further systems — COMET-Kiwi (wmt22), COMET-Kiwi (wmt23 XXL), MetricX, XCOMET, and GEMBA — were added subsequently, so the same segments carry six independent QE scores (Appendix D). The figures reported below are those of the first metric; extending the analysis to the others did not change the conclusions, and the per-metric results are reported in separate work. The object of critique here is segment-level QE scoring, whether based on fine-tuned encoders or on large language models, rather than any one implementation.

Scope of this report. A complete specification of the experiment — preprocessing, per-language-pair statistics, annotation procedure and annotator profiles, inter-annotator agreement, threshold definitions, and the resampling and cost-model formulas — is the subject of a separate paper and cannot be accommodated within the present format. The aim here is narrower: to establish whether a segment-level QE score can be relied upon to carry operational decisions. The underlying scores and MQM annotations are retained in full, and enquiries may be directed to the MQM Council at info@themqm.org.

Since TAUS EPIC QE reports a quality-oriented score, we treated lower TAUS scores as higher predicted error risk and evaluated the review order from the lowest TAUS score upward.

We then evaluated the score as a decision instrument, not merely as a correlation number. For each corpus/language-pair slice, we computed ROC AUC, AUPRC and AUPRC lift over random selection, precision, recall, specificity and F -scores at candidate thresholds, confusion matrices, triage-gain curves, calibration curves, Brier score relative

to a no-skill base-rate predictor, bootstrap confidence intervals, and permutation p-values.

The Revix export contained 186 snapshots. Each snapshot corresponds to a corpus/language-pair analysis under a specific threshold strategy and contains the resulting ranking, calibration, threshold, and workflow-utility statistics. Because several snapshots represented the same corpus/language-pair slice under different threshold strategies, the corpus-level analysis duplicated them into 72 unique dataset/language-pair slices. The threshold-specific analysis retained the full set of 186 snapshots, because the threshold choice is itself part of the operational decision. The saved threshold scenarios included maximum F_1 , maximum F_2 , maximum $F_{0.5}$, recall-at-least-90%, maximum Youden’s J , maximum triage gain, and manual or unspecified thresholds.

Aggregate finding. Table 1 shows that the QE score contains some statistical signal. The median AUC is above random for EC and EP, and many p-values are significant because the number of segments is large. However, the effect size is modest. The median EC AUC is only 0.591 and the median EP AUC is only 0.631. These values are compatible with weak-to-moderate ranking, not with reliable production classification.

The calibration results were even poorer. In every calibrated unique slice in the export, the Brier score was worse than the no-skill base-rate baseline; the Brier/no-skill ratio was above 1.0 for all 70 slices where calibration statistics were reported. This means that the score should not be interpreted as a probability-like estimate of translation quality or error risk. Even when the score is somewhat correlated with the presence of errors, it is not calibrated enough to support an absolute quality interpretation. This directly undermines the idea that the output is a general *translation quality score*.

Corpus-level behavior. The DFKI corpus illustrates why a statistically positive signal is not necessarily useful. Its error base rate was approximately 90–96%. In such a corpus, binary error/no-error triage is almost meaningless because nearly every segment is erroneous. The strongest AUC in the export occurred in this high-error regime, but that does not imply useful pass/fail triage. It means only that the score can sometimes rank very bad material ahead of less bad material in a corpus where almost everything already requires attention.

The EC corpus represents the opposite case. Its

Corpus	Slices	N	Base	AUC	Prec.@ F_1	Rec.@ F_1	Brier/no-skill
DFKI	2	2,543	93.3%	0.679	93.5%	100.0%	11.139
EC	20	45,504	6.5%	0.591	9.4%	85.8%	1.756
EP	23	56,715	32.5%	0.631	40.9%	95.5%	1.404

Table 1: Aggregate Revix results by primary corpus, for the TAUS EPIC QE scores. Base rate is the share of segments with at least one MQM error. Precision and recall are reported at the threshold that maximizes F_1 in each slice. A Brier/no-skill ratio above 1.0 means that the QE score is worse calibrated than a constant predictor that always returns the corpus base rate.

median base rate for any error was only 6.5%. In such a low-error regime, even a non-random ranker struggles to produce useful precision. For example, among the better EC slices, English–Greek reached AUC 0.692 but only 9.5% precision at peak F_1 ; English–Spanish reached AUC 0.662 but only 11.1% precision; English–Italian reached AUC 0.626 but only 7.3% precision. These results do not yield operationally persuasive targeted review lists. They can enrich an audit sample relative to random selection, but most flagged segments are still clean. For EC-like production-quality files, the evidence supports at most audit enrichment after local validation, not autonomous triage.

The EP corpus is the one where the QE score looks most plausible, because the base error rate is higher. The median base rate for any-error was 32.5%. The selected language pairs had moderate rank-order behavior: English–Czech reached AUC 0.695 with 48.2% precision at peak F_1 ; English–Italian reached AUC 0.673 with 35.4% precision; English–Polish reached AUC 0.670 with 42.0% precision; English–Dutch reached AUC 0.665 with 54.3% precision; English–Spanish reached AUC 0.664 with 44.5% precision. These figures show that the model can sometimes place more erroneous segments earlier for review. However, the weaker EP slices remained close to random for practical purposes: Irish, Latvian, Swedish, Estonian, Slovenian, and Hungarian all had AUCs below 0.60. Even in the best EP cases, the evidence can support the prioritization of segments for review, but not a safe threshold for bypassing the review.

Threshold behavior and triage. A central empirical result is that the operational verdict strongly depends on the threshold policy. Table 2 summarizes two important threshold strategies. The maximum- F_2 strategy emphasizes recall, while the maximum yield strategy emphasizes savings in the review effort.

The trade-off is clear. High recall usually requires flagging most of the segments in a file. In

EP data, the maximum- F_2 threshold catches almost all errors, but it flags a median of 80.8% of segments and has a negative median saving proxy (RCR-MEC, Modeled review-cost reduction, excluding missed-error cost). In EC data, the same strategy indicates a median of 70.6% of segments while producing only 9.1% precision. In contrast, the maximum-triage-gain strategy saves review effort but captures only about 60% of errors in both datasets, leaving a substantial share of errors in the unreviewed tail.

We also applied a stricter operational classification screen: a threshold would have to review at most 50% of the file, catch at least 80% of actual errors, and make the flagged set at least $1.5\times$ richer in errors than random selection. No saved threshold scenario in the export satisfied all three conditions.

The key operational finding is: If triage is defined as materially reducing review effort while catching the vast majority of errors, the data do not support segment-level QE as a safe triage mechanism.

Ranking is not triage. Ranking and triage are different questions: a ranker orders segments for a reviewer whose capacity is limited, whereas a triage gate decides which segments may bypass review altogether. Segment-level QE gives a weak answer to the first and no reliable answer to the second.

A QE model may produce a statistically useful ranking while still failing to support safe operational decisions under a specific threshold strategy, so the same score distribution can lead to different conclusions depending on the evaluation objective, as illustrated in Figure 1.

This distinction explains why the same corpus/language-pair slice could receive a “useful ranker” verdict under one threshold policy and a negative workflow verdict under another. Among the 72 unique slices, 50 had more than one saved verdict. Usability is therefore not an intrinsic property of the QE model: it is a property of the model,

Corpus	Threshold strategy	Snapshots	Median t	Flagged share	Precision	Recall	RCR-MEC
EC	Max F_2	20	0.908	70.6%	9.1%	93.7%	6.9%
EC	Max triage gain	18	0.878	42.7%	11.3%	60.5%	44.4%
EP	Max F_2	19	0.951	80.8%	39.8%	99.0%	-4.6%
EP	Max triage gain	23	0.891	42.0%	45.4%	60.4%	39.3%

Table 2: Threshold behavior for EC and EP, for the TAUS EPIC QE scores. RCR-MEC stands for "Modeled review-cost reduction, excluding missed-error cost": the value estimates the reduction in QE-plus-human-review expenditure relative to full human review. It excludes the operational, corrective, and risk costs of errors remaining in unreviewed segments. Under the maximum-triage-gain condition, approximately 40% of error-containing segments were not captured.

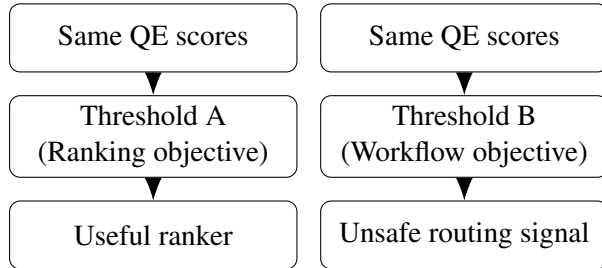


Figure 1: Identical QE scores can lead to different conclusions depending on the evaluation objective. A threshold selected to maximize ranking performance may yield a useful ranker, whereas a threshold selected to maximize workflow utility may reveal that the same scores are unsuitable as a routing signal. This illustrates that ranking quality and routing utility are not equivalent.

the corpus, the language pair, the error base rate, the threshold, and the relative cost of different translation mistakes. Any claim that a QE model is “usable” is insufficient unless it specifies the operating condition.

Non-portability across corpus conditions. The same target language did not behave consistently across EC and EP datasets. Across the 20 common target languages, the correlation between EC AUC and EP AUC was only 0.379. For English–Czech, AUC was 0.583 in EC but 0.695 in EP. For English–Swedish, AUC was 0.488 in EC and 0.586 in EP. In contrast, English–Portuguese and English–Greek were stronger in EC than in EP. The practical precision changed even more sharply because the base rates differed: a modest AUC in EC often produced a very low absolute error rate in the flagged set, while a similar AUC in EP produced a more persuasive targeted review list because the underlying error rate was much higher.

This matters for deployment: a threshold or language-pair conclusion reached on one dataset cannot be assumed to transfer to another file. Domain, data source, MT system, editorial status, and

expected error base rate all affect whether the QE score has any operational value.

What the results do and do not support. Read in isolation, the better EP slices might seem to license using the score to prioritize review. That reading draws the target around the arrow: the signal is identified after the fact, by selecting the corpus, language pair and threshold in which it appears, and it does not survive that selection. Across the twenty common target languages the correlation between EC and EP AUC is only 0.379, and the same slice receives opposite workflow verdicts under different threshold policies. A large corpus also makes a weak AUC statistically significant without making it useful for process control, and the scores are not calibrated, so they cannot be read as error probabilities. Thresholds that catch most errors flag most of the file, while thresholds that save review effort miss many. The apparent saving, moreover, is computed without the cost of the errors left unreviewed: RCR-MEC excludes precisely that cost, and about 40% of error-containing segments escape under the maximum-triage-gain condition. A prioritization gain that is neither reproducible across datasets nor priced against the errors it misses is not yet an operational result.

Production implication. The operational implication is that a segment-level QE tool cannot be used as a standalone routing, release, or review-bypass mechanism on a new file merely because it produced useful statistics elsewhere. Before using it in production, local validation on the material with the same language pair, domain, MT or AI translation source, editorial condition, and comparable expected error rate would be needed. A defensible workflow would sort segments worst-first, retain random or stratified sampling from the supposedly good tail, and validate against human MQM annotation that the top score bands actually contain a materially higher error density than ran-

dom selection. Without such validation, the score is not a safe decision variable.

Three considerations indicate that this reflects a property of segment-level scoring rather than of one vendor’s model. The argument of Section 3 is structural, and changing the scorer does not remove the constraint it identifies. The same failure mode is visible with a different scorer on different data: the distribution in Figure 2 was produced with a COMET QE model. And repeating the analysis on the five further metrics scored on this material (Appendix D) yielded the same qualitative outcome.

Segment-level QE may contain linguistic or distributional information learned from training data, but this information is not equivalent to information about translation quality: it is a corpus-dependent proxy signal. What matters is not the score itself but whether it is discriminative, calibrated, threshold-stable, economically meaningful, and safe under the conditions of deployment.

5 Conclusion

The cumulative evidence from the literature reviewed and from the present analysis suggests that QE technology remains unstable across multiple dimensions, which raises serious concerns about its suitability for real-world applications.

The persistent failure of QE systems to handle phenomena such as homophones, together with the findings on length bias, suggests that current architectures rely on superficial statistical patterns rather than genuine semantic evaluation, and the practical utility of QE in professional workflows is far from established. Taken with domain generalization failures, annotation inconsistencies, and the difficulty of fine-grained error classification, this indicates that QE systems still lack a robust and portable understanding of translation quality; more data and more sophisticated architectures have not yet resolved these structural issues.

Future progress will require explicit attention to the statistical uncertainty inherent in both human and AI judgment. As Sarti et al. (2025b) demonstrate, single-pass evaluations yield unstable metric rankings. Whether assessment is performed by human annotators, automated models, or hybrid systems, evaluator variance must be recognized, measured, and reported—through multi-evaluator sampling, prior uncertainty calibration, or probabilistic scoring.

MQM-based approaches differ from QE in that

they explicitly model multiple dimensions of translation quality and allow contextual factors to be taken into account, which makes them better suited to its multi-faceted, context-dependent nature. MQM is a framework rather than a procedure: it specifies error typologies and severities, quality specifications, a scoring model and sampling requirements, and it treats human perception as the gold standard because translation quality is a property of how a text is received. It does not require that every judgment be made by a human. Throughput objections therefore apply to particular implementations of MQM, not to the framework, which an automated evaluator calibrated against human-based MQM can in principle satisfy as fully as a human one.

Rather than continuing to prioritize fully automated QE as a standalone approach, future research may benefit from methods that more explicitly account for the complexity of human language and judgment. The available evidence suggests that segment-level QE scores are not sufficiently reliable to serve as standalone signals for routing decisions or release criteria, and that whatever prioritization value they carry is condition-dependent and must be established by local validation rather than assumed. Addressing these limitations may require moving beyond isolated, segment-level prediction toward scorers that can observe the contextual term *c* at all — document- and paragraph-level models with multi-segment context windows — and toward automated evaluation grounded in structured, multi-dimensional assessment such as MQM, calibrated against human-based MQM and recalibrated as models and technology drift. Human annotation is the reference that automation must be held to, not a substitute for it.

Disclaimer

The views expressed in this paper are solely those of the authors and do not represent the official position of the European Parliament or its Directorate-General for Translation and Clear Language.

Acknowledgments

We thank Fred Bane, Director of Data Science at TransPerfect for valuable comments on this paper.

References

Fernando Alva-Manchego, Lucia Specia, Sara Szoc, Tom Vanallemeersch, and Heidi Depraetere. 2021.

- Validating quality estimation in a computer-aided translation workflow: Speed, cost and quality trade-off. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*.
- Hanna Béchara, Constantin Orăsan, Carla Parra Escartín, Marcos Zampieri, and Will Lowe. 2021. [The role of machine translation quality estimation in the post-editing workflow](#). *Informatics*, 8(3).
- John Cunnison Catford. 1965. *A linguistic theory of translation*, volume 31. Oxford university press London.
- Victor Chahuneau, Eva Schlinger, Noah A Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687.
- Tu Anh Dinh and Jan Niehues. 2023. [Perturbation-based QE: An Explainable, Unsupervised Word-level Quality Estimation Method for Blackbox Machine Translation](#). In *Proceedings of Machine Translation Summit XIX*.
- Barbara Dragsted. 2004. *Segmentation in Translation and Translation Memory Systems: An Empirical Investigation of Cognitive Segmentation*. Ph.D. thesis, Copenhagen Business School.
- Xiang Geng, Zhejian Lai, Jiajun Chen, Hao Yang, and Shujian Huang. 2025. [Alleviating distribution shift in synthetic data for machine translation quality estimation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Serge Gladkoff. 2025. Hard fact: Automatic quality prediction based on neural models does not work at the segment level. Logrus Global explainer publication. Industry report, accessed June 2026.
- Serge Gladkoff, Irina Sorokina, Lifeng Han, and Anastasia Alekseeva. 2022. [Measuring Uncertainty in Translation Quality Evaluation \(TQE\)](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.
- Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. [Uncertainty-aware machine translation evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. [xcomet: Transparent machine translation evaluation through fine-grained error detection](#). *Preprint*, arXiv:2310.10482.
- Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, 6:1211564.
- Fei Huang, Jian-Ming Xu, Abraham Ittycheriah, and Salim Roukos. 2014. [Adaptive HTER Estimation for Document-Specific MT Post-Editing](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–870.
- Roman Jakobson. 1959. On linguistic aspects of translation. In Reuben A. Brower, editor, *On Translation*, pages 232–239. Harvard University Press, Cambridge, MA.
- Dongjun Jung, Chanjun Park, Sugyeong Eo, and Heuseok Lim. 2023. [Enhancing machine translation quality estimation via fine-grained error analysis and large language model](#). *Mathematics*, 11(19).
- Mehmet Kocyigit, James Lee, and Derry Wijaya. 2022. [Better quality estimation for low resource corpus mining](#). In *Findings of the Association for Computational Linguistics: ACL 2022*.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. [Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.
- Dongjun Lee, Jaehui Ahn, Hwanhee Park, and Jihun Jo. 2021. [Intellicat: Intelligent machine translation post-editing with quality estimation and translation suggestion](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 11–19.
- Jiahui Liang and Lifeng Han. 2026. [MetaHOPE: A metaphor-oriented evaluation framework for analysing mt and llm translation errors](#). In *Proceedings of the 9th International Conference on Natural Language and Speech Processing (ICNLSP 2026)*. Accepted for publication. Preprint available at arXiv:2607.00848.
- Siqi Liu, Guangrong Dai, and Dechao Li. 2025. [Introducing quality estimation to machine translation post-editing workflow: An empirical study on its usefulness](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 485–495, Geneva, Switzerland. European Association for Machine Translation.
- Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. [The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control](#). In *Proceedings of the 16th*

- Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.
- Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. [Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.
- Shenbin Qian, Constantin Orăsan, Diptesh Kanojia, and Félix do Carmo. 2025. [Automatically generating chinese homophone words to probe machine translation estimation systems](#). *Preprint*, arXiv:2503.16158.
- Tharindu Ranasinghe, Constantin Orăsan, and Ruslan Mitkov. 2020. Intelligent translation memory matching and retrieval with sentence encoders. *arXiv preprint arXiv:2004.12894*.
- Andy Rosenbaum, Assaf Siani, and Ilan Kernerman. 2026. [MTQE.en-he: Machine translation quality estimation for English-Hebrew](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 559–569, Rabat, Morocco. Association for Computational Linguistics.
- Gabriele Sarti, Vilém Zouhar, Grzegorz Chrupała, Ana Guerberof-Arenas, Malvina Nissim, and Arianna Bisazza. 2025a. [QE4PE: Word-level Quality Estimation for Human Post-Editing](#). *Transactions of the Association for Computational Linguistics*.
- Gabriele Sarti, Vilém Zouhar, Malvina Nissim, and Arianna Bisazza. 2025b. [Unsupervised word-level quality estimation for machine translation through the lens of annotators \(dis\)agreement](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18320–18337.
- Carolina Scarton, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia. 2019. [Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality](#). In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.
- Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, Frédéric Blain, Eva Vanmassenhove, Mirella De Sisto, Chris Emmery, and Pieter Spronck. 2023. [Tailoring domain adaptation for machine translation quality estimation](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.
- Assaf Siani, Anna Kernerman, and Ilan Kernerman. 2026. [Semi-synthetic parallel data for translation quality estimation: A case study of dataset building for an under-resourced language pair](#). *Preprint*, arXiv:2603.11743.
- Maria Carmen Staiano, Lifeng Han, Johanna Monti, and Francesca Chiusaroli. 2025. [ITALERT: Assessing the quality of LLMs and NMT in translating Italian emergency response text](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 566–577, Geneva, Switzerland. European Association for Machine Translation.
- Marco Turchi, Matteo Negri, and Marcello Federico. 2015. [MT Quality Estimation for Computer-assisted Translation: Does it Really Help?](#) In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 530–535.
- Zhen Yang, Fandong Meng, Yuanmeng Yan, and Jie Zhou. 2023. Rethinking the word-level quality estimation for machine translation from human judgement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2012–2025.
- Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and André Martins. 2025. [Watching the watchers: Exposing gender disparities in machine translation quality estimation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Anna Zaretskaya, José Conceição, and Frederick Bane. 2020. [Estimation vs metrics: is QE useful for MT model selection?](#) In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 339–346, Lisboa, Portugal. European Association for Machine Translation.
- Yilin Zhang, Wenda Xu, Zhongtao Liu, Tetsuji Nakagawa, and Markus Freitag. 2025. [Penalizing length: Uncovering systematic bias in quality estimation metrics](#). *Preprint*, arXiv:2510.22028.

A Automatic QE Scores versus Actual Error Distributions

Figure 2 shows automatic quality prediction scores against the actual distribution of annotated errors, from the industrial case study reported by Gladkoff (2025).

B Linguistic and Terminology Examples

Linguistic Example: Consider the Spanish source segment "Tiene razón" and its translation into target languages that require grammatical gender marking for pronouns used to resolve the autonomous Spanish verb forms. Whether a particular translation is correct depends on discourse context establishing the referent's gender ("he/she is right"), communicative register (if the register is formal, this could mean "you are right"), or whether it refers to an entity (e.g. the Government, "it is right"). Without access to sufficient contextual information, multiple distinct translations may each be plausible, and no sentence-level QE model can determine which variant is appropriate based on the segment alone. More generally, identical source–translation segment pairs may receive different quality judgments depending on contextual variables external to the segment itself, illustrating that sentence-level translation quality is not uniquely determined by isolated segment pairs.

Terminology/Domain Example: Consider the source segment "Apply the patch." In a software documentation context, patch may refer to a software update, whereas in a medical context, it may refer to a transdermal medication patch. A translation that is correct in one domain may be incorrect in another, even if the isolated segment pair remains identical. Without domain context, sentence-level QE cannot reliably determine translation correctness.

Therefore, without domain context, segment-level QE is forced to evaluate translations against moving targets.

C Limited Assistance, No Routing Validation for Segment-Level QE

The literature on sentence-level QE can be broadly divided into two distinct lines of inquiry: (i) whether QE information can assist human translators in constrained MT post-editing scenarios, and (ii) whether sentence-level QE scores can serve as reliable standalone signals for production routing decisions, such as determining which segments

may bypass human review. These two questions are frequently conflated, although existing evidence does not support treating them as equivalent.

On the first question, the evidence is limited and mixed. Some studies report partial or conditional benefits when QE information is shown to human post-editors. Huang et al. (2014) report productivity gains in a document-specific post-editing setting. Turchi et al. (2015) find only slight gains overall, with statistical significance limited to restricted conditions of segment length and MT quality. Béchara et al. (2021) report lower average post-editing time and effort when MT QE information was provided, but their study is relatively small. More recently, Liu et al. (2025) concluded that empirical evidence for the usefulness of sentence-level QE remains limited and inconsistent. Importantly, these studies are about human-in-the-loop assistance, not autonomous acceptance, rejection, or routing of segments. Integrated CAT-tool studies such as IntelliCAT (Lee et al., 2021) report productivity improvements, but because sentence-level QE is bundled with other assistive features, the contribution of QE alone cannot be isolated.

To our knowledge, the main published attempt to validate sentence-level QE as an explicit routing mechanism is the APE-QUEST case study Alva-Manchego et al. (2021). This study reports that, in a narrowly defined public-administration workflow for English–Dutch and English–French, thresholded QE scores could reduce time and cost while preserving roughly similar end-user acceptability. However, these results should not be interpreted as a general validation of sentence-level routing. It is a single case study restricted to two language pairs, one domain, and two specific use cases; it depends on threshold tuning for each scenario; and the authors themselves show that gains are larger with oracle scores than with predicted ones.

By contrast, the counter-evidence is more directly relevant to the production-routing claim. Zaretskaya et al. (2020) find that an industry-level QE system was not reliable enough for the selection of MT engines when candidate systems had similar quality. Glushkova et al. (2021) argue that segment-level point estimates provide limited information and are often unreliable because they are trained on noisy, biased, and scarce human judgments. Most recently, the WMT25 Shared Task confirmed that segment level prediction remains difficult, reference-based baselines still outperform LLMs at the segment level, accurate error

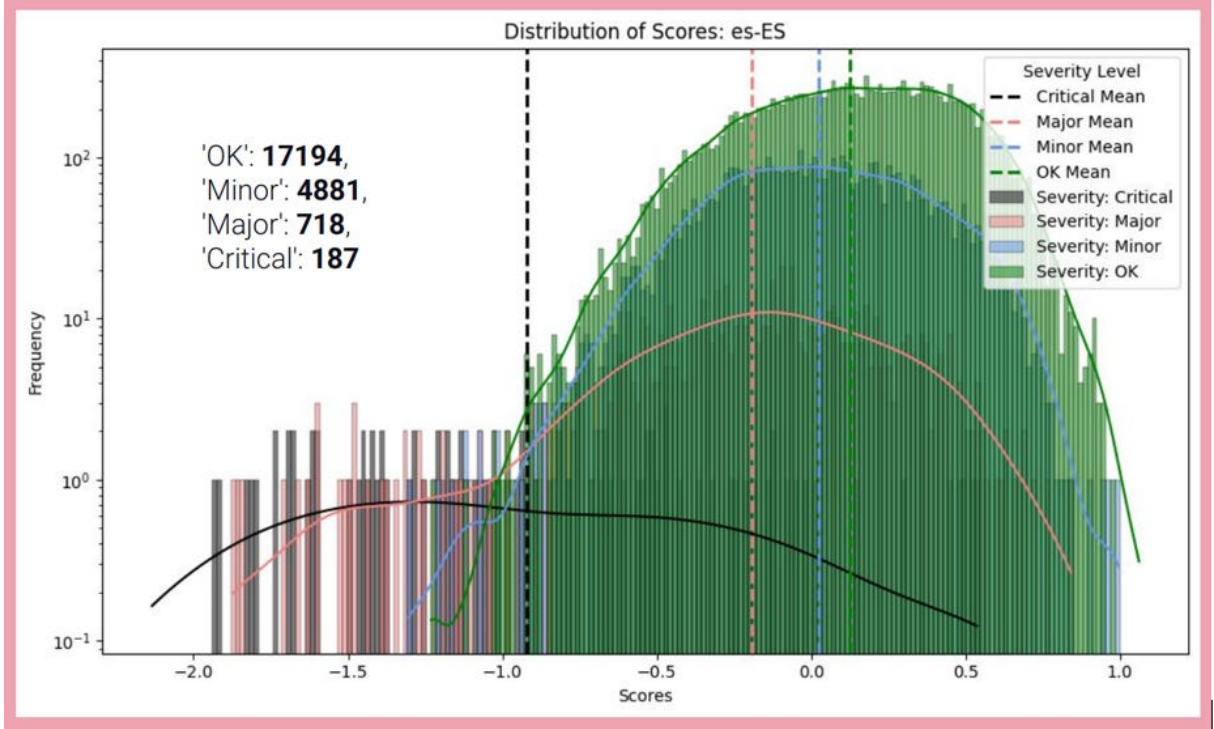


Figure 2: Automatic translation quality prediction scores vs. actual error distributions. (COMET model wmt20-comet-qe-da). Note that Y-axis is in log scale.

detection remains challenging, and even quality-informed minimal correction is difficult across diverse languages (Lavie et al., 2025). The strongest production-oriented negative evidence is the large-scale industry case study reported by Gladkoff (2025). In that work, severe and critical errors overlap broadly across the sentence-level QE score range, so that no reliable threshold can separate safe-to-pass segments from those requiring review. If serious errors are distributed across the full score continuum, then the score fails not only as a cut-off variable, but also as a ranking variable to prioritize the review effort (see Figure 2). Consequently, the current literature does not provide robust and replicated support for the use of raw segment-level QE scores as standalone production routing criteria. At most, it provides limited and mixed evidence that QE information may sometimes assist human post-editors in narrowly defined workflows. That is a materially weaker claim, and it does not support broader claims in the literature and practitioner discourse that segment-level QE is ready for generalized no-touch routing in production.

D Metric Provenance and Coverage

Quality-estimation scores for the annotated material were obtained from six systems and

are stored per segment together with a metric identifier and, where the upstream interface exposes one, a model version. The identifiers are `taus_qe`, `wmt22-cometkiwi-da`, `wmt23-cometkiwi-da-xxl`, `metricx`, `xcomet` and `gemba`. Table 3 reports the model or checkpoint and the number of scored rows for each.

Metric	Version or checkpoint	Rows
TAUS EPIC AIQE	2.0.0	120,619
COMET-Kiwi	wmt22-cometkiwi-da	141,641
COMET-Kiwi	wmt23-cometkiwi-da-xxl	141,641
MetricX	reference-free	9,588
XCOMET	recorded per row	9,588
GEMBA	recorded per row	7,760

Table 3: Provenance and coverage of the six quality-estimation metrics scored on the annotated material. Full checkpoint identifiers are given in the text. Row counts are all scored rows held in the analysis database; the analysis reported in Section 4 uses a 104,762-segment subset drawn from the three MQM-annotated corpora described there.

TAUS EPIC AIQE was called against the commercial endpoint <https://api.taus.net/2.0/estimate/>. TAUS serves the current production model for each language pair and does not permit pinning to an earlier revision, but the metric identifier and version returned by the interface are recorded for every segment. Across our data every

TAUS-scored segment was returned under a single identifier and version, so the corpus is directly comparable on this metric. The two COMET-Kiwi scores were computed against the publicly released checkpoints Unbabel/wmt22-cometkiwi-da and Unbabel/wmt23-cometkiwi-da-xxl.

MetricX, XCOMET, GEMBA and both COMET-Kiwi checkpoints were computed outside Revix and have been imported as pre-computed score columns. Their exact checkpoint identifiers are therefore attested by the parties that computed them rather than by Revix records. A further 1,940 segments carry a pre-computed TAUS score that arrived without an accompanying version tag; these were excluded from version-sensitive TAUS analyses but retained for cross-metric comparison. Coverage is uneven: TAUS EPIC AIQE and both COMET-Kiwi checkpoints span the full scored corpus, whereas MetricX, XCOMET and GEMBA were run on smaller subsets. The scores and the corresponding MQM annotations are retained in full, and enquiries concerning access may be directed to the MQM Council.