Multilingual Machine Translation Quality Estimation: Evaluating a Production-Grade MTQE Framework

Benchmarking Safe Publication decisions, Automatic Post-Editing, and Human review across 16 Language pairs

Technical Whitepaper

Pangeanic B.I Europa S.L
MTQE v2
Yash Namdeo Dhobale • Manuel Herranz Pérez • Marina Albert Girona

August 2026

INDEX

Executive Summary

Enterprise translation has moved beyond the question of whether AI can generate fluent multilingual text. The operational question is whether an organization can trust, govern, and deploy that output at scale. General-purpose models make translation abundant, but they do not by themselves provide the quality thresholds, decision logic, cost control, or accountability required for production. A reliable enterprise workflow therefore needs a specialized quality layer that evaluates every source–translation pair before publication.

This whitepaper presents **Pangeanic MTQE v2**, a multilingual machine translation quality estimation system built for that role. Rather than returning an abstract quality score alone, it converts translation evidence into an operational decision: accept a translation that meets the publication threshold, repair a recoverable translation through automatic post-editing (APE), or escalate the remaining case for human post-editing (HPE). This approach applies the broader enterprise-AI principle that clearly defined tasks are best served by specialized, governable models integrated with organizational data, policies, and human oversight.

Executive benchmark snapshot		
Dimension	**Result**	**Operational interpretation**
Incorrect segments evaluated	6,006	Weighted multilingual benchmark population used to assess unsafe acceptance and downstream routing
Incorrect segments accepted	65	Incorrect translations that passed the MTQE quality gate
Weighted correct-rejection accuracy	98.9%	The system blocked nearly all known incorrect translations from direct acceptance
False acceptance rate	1.1%	Primary residual publication-risk indicator
APE recovery rate	67.6%	More than two-thirds of incorrect segments were automatically repaired, making post-editing the principal downstream workload
Weighted mean latency per segment	0.49 seconds	Supports high-throughput, near-real-time quality routing

For decision-makers, the implication is a shift from universal human inspection to governed exception handling. Human expertise remains essential, but it is concentrated on the minority of cases that cannot be accepted or repaired automatically. The organization retains control over thresholds, deployment, language coverage, and escalation policy while gaining a measurable route to faster publishing, lower review demand, and more consistent multilingual risk management.

1. Background and Motivation

Machine translation is now remarkably good. In many production settings, typical output may be approximately 90% accurate and often appears indistinguishable from fully correct translation. That apparent fluency creates a dangerous confidence gap: the remaining errors are difficult to see, yet they carry a disproportionate share of business risk. In legal, medical, financial, and brand-sensitive content, a mistranslated obligation, dosage, amount, name, claim, or instruction can make an otherwise fluent segment unusable. The last 10% matters most.

At the same time, translation volume has expanded to millions of segments per day. Reviewing every segment manually is prohibitively expensive, slows publication, and consumes scarce linguistic expertise confirming text that is already correct. Quality also varies across languages, models, domains, and content types, while modern AI systems can still omit information, change meaning, or hallucinate plausible details. The operational challenge is therefore not to review everything; it is to identify the small, high-risk subset that genuinely requires attention.

Conventional machine translation pipelines generate output but do not answer the questions an enterprise must resolve before trusting it: How reliable is this translation? Should it be accepted? Which sentence needs attention? Why is it considered problematic? Without a dedicated confidence layer, organizations must choose between two unsatisfactory options - trust output without sufficient evidence or route nearly everything through costly human review.

2. Existing Approaches

Traditional linguistic checks identify formatting, spelling, grammar, terminology, and rule-based violations. Reference-based metrics such as BLEU quantify lexical overlap with one or more references. Learned metrics such as COMET and MetricX aim to align more closely with human quality judgments. Human review remains the final authority in high-risk workflows. Each approach is valuable, but none automatically converts evidence into the specific operational action required for every segment.

Approach comparison			
Approach	**Primary output**	**Strength**	**Operational limitation**
Rules and terminology checks	Pass/fail flags	Transparent and targeted	Limited semantic coverage
BLEU / lexical metrics	Reference-overlap score	Fast and reproducible	Requires references; valid paraphrases may be penalized
COMET / learned metrics	Quality score	Stronger semantic modeling	A score is not a calibrated publication decision
Human review	Expert judgment	Contextual and adaptable	Expensive, slow, and capacity constrained

3. Benchmark Methodology

3.1 Overview

The objective of this benchmark is to evaluate the performance of Pangeanic MTQE v2 as a production-oriented Machine Translation Quality Estimation (MTQE) system across multiple languages and translation scenarios.

Unlike traditional machine translation evaluation metrics that compare machine translations against reference translations, MTQE operates in a reference-free setting. Each evaluation requires only the original source segment and its translated target segment, allowing quality estimation to be performed directly on production translations without human references.

For every evaluated segment, MTQE predicts the overall translation quality and determines whether the translation is suitable for publication or requires further intervention.

3.2 Evaluation Workflow

Each benchmark segment was evaluated independently using the MTQE v2 API.

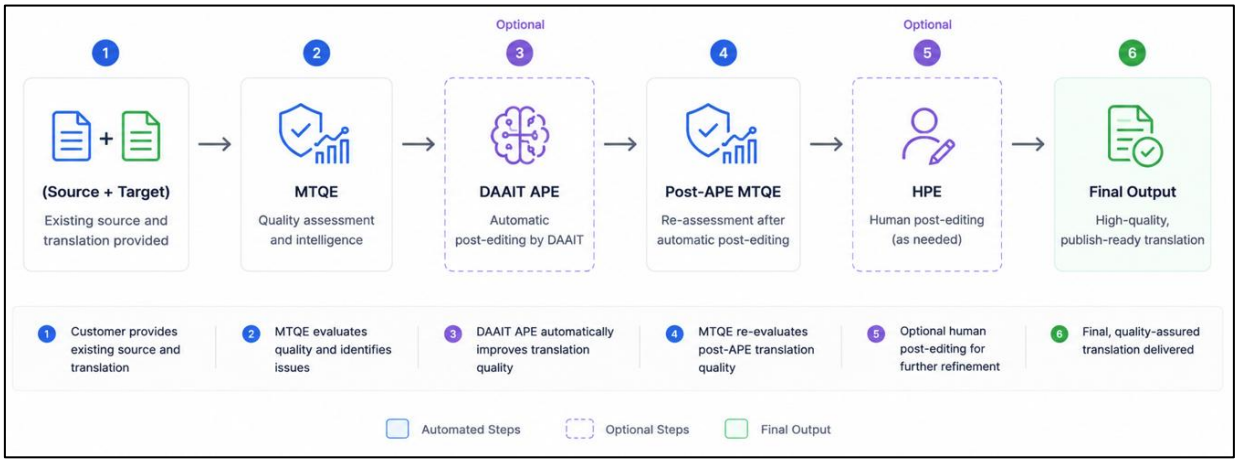
For every sample, the benchmark provided:

- Source language
- Target language
- Source segment
- Candidate translation

No reference translations or additional annotations were supplied to the system during inference.

The MTQE model analyzed each source–translation pair and generated:

- A quality score ranging from 0 to 100.
- A natural-language explanation describing detected translation issues.
- An optional Automatic Post-Editing (APE) output for translations below the configured quality threshold using Pangeanic's proprietary Deep Adaptive AI Translation Platform (DAAIT)
- A re-evaluated quality score after post-editing when APE was enabled.
- Low Quality segments were further classified for Human Post-Editing



This workflow mirrors the intended production deployment of the system, where translations are evaluated immediately after generation before publication or further review.

3.3 Decision Strategy

For benchmarking purposes, translation quality was interpreted using the operational workflow implemented by MTQE. Each translation followed the decision process illustrated below.

Benchmark decision process - For each source–translation pair, the fine-tuned MTQE model jointly analyzes the predicted quality score and the nature of the detected translation error. Based on this combined assessment, the model determines whether the segment is acceptable or should be classified for post-editing.

3.4 Automatic Post-Editing

To simulate a production-quality localization workflow, Automatic Post-Editing (APE) was enabled during benchmarking using **Deep Adaptive AI Translation (DAAIT)**, Pangeanic's proprietary adaptive translation model. DAAIT is designed not simply to rewrite low-quality output, but to produce an improved translation that remains grounded in approved terminology, prior translations, domain knowledge, and operational quality requirements.

Within the benchmark, whenever a translation failed the initial MTQE evaluation, DAAIT generated a post-edited candidate and returned it to MTQE for a second quality assessment. This separation of responsibilities preserved the role of MTQE as the quality gate while using DAAIT as the corrective layer. The sequence also

reflects governed automation: translation assets, adaptation behavior, correction, re-scoring, and escalation can be orchestrated as an auditable workflow rather than an unbounded generative step.

This two-stage evaluation enabled the benchmark to distinguish between -

- Translations accepted immediately.
- Translations successfully recovered through automatic post-editing.
- Translations that ultimately required human intervention.

Rather than measuring MTQE as a standalone classifier, this approach evaluates its effectiveness as an operational quality gate within a complete localization workflow.

3.5 API Configuration

All benchmark experiments were executed using the production implementation of the Pangeanic MTQE v2 REST API. Three API modes are available:

- Single-segment synchronous evaluation
- Batch evaluation
- Asynchronous file evaluation

The benchmarking experiments primarily used the file-processing endpoint, allowing complete benchmark datasets to be evaluated under identical configuration settings while preserving consistent language parameters and scoring behavior.

For all experiments:

- Source and target languages were explicitly specified.
- Identical scoring configuration was used across all language pairs.
- Automatic Post-Editing was enabled.
- Translation memories and glossaries were disabled unless explicitly stated.

3.6 Recorded Outputs

For every benchmark segment, the following outputs were recorded:

Recorded MTQE and post-editing outputs		
API field	**Report label**	**Description**
score	Initial MTQE Score	Initial reference-free quality score on a 0–100 scale.
explanation	Initial Explanation	Natural-language description of the translation issue detected in the original candidate. This field may be empty when no material issue is identified.
ape_needed	APE Required	Boolean routing flag indicating whether the original candidate was classified for Automatic Post-Editing.
target_APE	APE Translation	Post-edited target segment generated by DAAIT when APE is required.
score_APE	Post-APE MTQE Score	Quality score assigned after the DAAIT output is re-evaluated by MTQE.
explanation_ape	Post-APE Explanation	Natural-language assessment of any residual issue detected after post-editing. The field may be blank when the revised translation is considered fully acceptable.
hpe_needed	Human Post-Editing Required	Boolean routing flag indicating whether the segment requires human post-editing after the MTQE and APE stages.

For analysis, the final workflow outcome was derived from the routing fields. A segment with ape_needed = false was accepted without automatic correction. A segment with ape_needed = true was processed by DAAIT and re-scored by MTQE. After that stage, hpe_needed = false indicated successful automatic recovery, while hpe_needed = true indicated escalation to Human Post-Editing. These fields formed the basis of the false acceptance, rejection accuracy, APE recovery, HPE, and multilingual performance analyses.

4. Benchmark Dataset

4.1 ACES: Translation Accuracy Challenge Sets

The primary benchmark used in this study is **ACES (Translation Accuracy Challenge Sets)**, a multilingual evaluation dataset designed to test whether machine translation evaluation systems can detect translation accuracy errors at the segment level. Introduced by Amrhein, Moghe, and Guillou in 2022 and extended through subsequent large-scale analysis, ACES addresses a limitation of conventional benchmarks: aggregate correlation with human judgments can summarize overall metric quality, but it reveals little about reliability on specific error types.

ACES uses a contrastive evaluation paradigm. Each benchmark instance contains:

- The original source sentence.
- A higher-quality ("good") translation.
- A minimally contrasted, deliberately degraded ("incorrect") translation containing a controlled translation error.
- A human reference translation.
- The corresponding language pair.
- An annotation identifying the linguistic phenomenon represented by the error.

This structure makes it possible to evaluate whether a quality estimation system distinguishes an acceptable translation from a closely related alternative containing a specific semantic or linguistic defect. Because the two candidates are intentionally similar, success depends on identifying the controlled error rather than rewarding general fluency or superficial overlap.

4.2 Dataset Scale and Linguistic Coverage

The benchmark contains 36,476 examples spanning 146 language pairs and 68 translation phenomena. Its coverage includes high-, medium-, and lower-resource languages across multiple language families, writing systems, and translation directions. This breadth enables evaluation of multilingual robustness rather than performance on a narrow group of commonly tested language pairs.

The represented phenomena extend beyond lexical substitution and grammatical well-formedness. They include additions, omissions, mistranslations, named entities, numerical expressions, negation, word-sense disambiguation, agreement, punctuation, discourse consistency, anaphora, real-world knowledge, and other accuracy-related phenomena informed by the Multidimensional Quality Metrics framework. The collection therefore probes errors that may substantially change source meaning even when the target remains fluent.

4.3 Why ACES Was Selected

ACES was originally designed to evaluate machine translation metrics by testing whether a metric consistently assigns a higher score to the good translation than to the incorrect alternative. Large-scale evaluations across more than fifty metrics have shown substantial variation by phenomenon and no single metric that performs consistently across all error categories. The analyses also indicate that many methods rely too heavily on surface-level similarity and insufficiently model semantic correspondence with the source.

These properties make ACES particularly suitable for evaluating a production-oriented MTQE system. It intentionally stresses the evaluator with controlled failures that may appear fluent, permits phenomenon-level error analysis, and covers diverse languages and scripts. It therefore supports a more rigorous assessment of source-sensitive error detection and multilingual robustness than an aggregate-quality dataset alone.

4.4 Use of Known Incorrect Translations

ACES was selected primarily because it provides a large, clearly labeled collection of incorrect translations derived from known source segments. Each incorrect candidate contains an intentional translation error, creating an unambiguous test set of outputs that should not pass an operational quality gate. This makes the dataset especially useful for measuring the central safety objective of Pangeanic MTQE v2: identifying translations that are known to be wrong before they reach publication.

In this evaluation, each ACES incorrect translation was submitted independently with its corresponding source segment. The expected behavior was that MTQE would recognize the semantic or linguistic defect and classify the segment for

corrective action. Ideally, every intentionally incorrect translation would be detected. Any incorrect segment accepted without intervention was therefore treated as a false acceptance and as the most important residual quality risk.

This use of ACES focuses the benchmark on error-detection reliability rather than relative ranking between candidate translations. It enables direct measurement of how many known incorrect segments MTQE rejects, how many it routes to DAAIT for automatic correction, how many are ultimately escalated for human post-editing, and most critically how many incorrect translations remain undetected.

4.5 Language Pairs Evaluated

The benchmark evaluation used 16 language pairs selected from ACES. The set includes English as both a source and target language, direct non-English translation directions, European and Asian languages, and languages written in Latin, Arabic, Cyrillic, Chinese, Japanese, Korean, and Thai scripts. This mix was intended to test whether MTQE could identify known incorrect translations consistently across different language families, scripts, resource levels, and translation directions.

Language pairs included in the benchmark		
Source language	**Target language**	**Pair**
English	Spanish	en → es
Japanese	English	ja → en
German	Chinese	de → zh
Spanish	French	es → fr
English	Korean	en → ko
English	Russian	en → ru
Vietnamese	English	vi → en
Arabic	English	ar → en
Swedish	English	sw → en
Thai	English	th → en
Japanese	Korean	ja → ko
Spanish	Japanese	es → ja

Lithuanian	English	lt → en
Chinese	English	zh → en
Chinese	German	zh → de
Japanese	Spanish	ja → es

5. Benchmark Results

5.1 Overall Results

Across 6,006 known incorrect ACES segments, Pangeanic MTQE v2 accepted only 65, producing a weighted false acceptance rate of 1.1% and a correct-rejection rate of 98.9%. DAAIT automatically repaired 4,058 segments, equivalent to 67.6% of all incorrect inputs. The remaining 1,883 segments required Human Post-Editing, representing 31.4% of the benchmark. Total processing time was 2,944 seconds, or a weighted mean of 0.49 seconds per segment.

Overall results for known incorrect segments		
Measure	**Result**	**Interpretation**
Incorrect segments evaluated	6,006	Full multilingual test population
Incorrect segments accepted	65	False acceptances
False acceptance rate	1.1%	Primary residual publication risk
Correct-rejection rate	98.9%	Known incorrect segments blocked from direct acceptance
Fixed by DAAIT APE	4,058 (67.6%)	Incorrect segments recovered automatically
Human Post-Editing required	1,883 (31.4%)	Incorrect segments escalated after automated processing
Total processing time	2,944 seconds	Aggregate benchmark runtime
Weighted mean latency	0.49 seconds per segment	Suitable for high-throughput routing

The aggregate result shows that the system's principal contribution was not only detection but recovery. Nearly seven in ten incorrect translations were repaired automatically, while approximately three in ten were reserved for human intervention. The small false-acceptance population therefore becomes the most important group for detailed qualitative analysis.

5.2 Performance by Language Pair

Performance varied by language direction. Seven pairs recorded zero false acceptances: vi→en, ar→en, sw→en, th→en, ja→ko, lt→en, and zh→de. The lowest non-zero rate was 0.2% for en→ru. The largest false-acceptance concentration occurred in es→ja at 6.8%, followed by zh→en at 2.4% and ja→en at 1.8%. Because sample sizes differ substantially, both rates and counts should be considered.

Incorrect-segment performance by language pair								
Lang pair	**Total segments**	**Accepted**	**Fixed by APE**	**Need HPE**	**Total time (in seconds)**	**Avg. time per segment**	**Average score**	**% accuracy**
en-es	725	5	514	206	270	0,37	67,96	99,31%
ja-en	912	16	634	262	430	0,47	78,89	98,25%
de-zh	75	1	46	28	34	0,45	78,3	98,67%
es-fr	125	1	86	38	49	0,39	75,98	99,20%
en-ko	545	3	340	202	229	0,42	75,52	99,45%
en-ru	472	1	376	95	173	0,37	57,31	99,79%
vi-en	391	0	276	115	156	0,40	50,91	100,00%
ar-en	361	0	255	106	143	0,40	45,14	100,00%
sw-en	327	0	71	256	153	0,47	42,05	100,00%
th-en	299	0	204	95	127	0,42	49,89	100,00%
ja-ko	163	0	114	49	106	0,65	55,09	100,00%
es-ja	117	8	64	45	114	0,97	76,84	93,16%
lt-en	68	0	57	11	100	1,47	79,52	100,00%
zh-en	1209	29	883	297	435	0,36	71	97,60%
zh-de	150	0	99	51	179	1,19	53,99	100,00%
ja-es	67	1	39	27	77	1,15	76,05	98,51%

Automatic recovery was strongest for lt→en (83.8%) and en→ru (79.7%), whereas sw→en was a clear outlier at 21.7% APE recovery and 78.3% HPE. This pattern suggests that reliable rejection and successful automatic correction are distinct capabilities: the system rejected every known incorrect sw→en segment, but most still required human intervention. Runtime was below half a second per segment for 11 of the 16 pairs; the slowest averages were lt→en (1.47 seconds), zh→de (1.19 seconds), and ja→es (1.15 seconds).

5.3 False Acceptance Analysis

False acceptance is the benchmark's most consequential failure mode because it represents a known incorrect translation that was not routed for correction. The 65 false acceptances were concentrated rather than evenly distributed. zh→en contributed 29 cases (44.6% of all false acceptances), ja→en contributed 16 (24.6%), and es→ja contributed 8 (12.3%). Together, these three directions accounted for 81.5% of the total false-acceptance population.

Highest False-Accept Risk					
Pair	**Segments**	**Accepted**	**False accept rate**	**Correct rejection**	**Average score**
es→ja	117	8	6,8%	93,2%	76,8
zh→en	1.209	29	2,4%	97,6%	71,0
ja→en	912	16	1,8%	98,2%	78,9
ja→es	67	1	1,5%	98,5%	76,1
de→zh	75	1	1,3%	98,7%	78,3
es→fr	125	1	0,8%	99,2%	76,0
en→es	725	5	0,7%	99,3%	68,0
en→ko	545	3	0,6%	99,4%	75,5

Zero false-acceptance language pairs. Seven directions recorded no false acceptances and therefore achieved 100% correct rejection on the evaluated

incorrect segments: vi→en (391 segments), ar→en (361), sw→en (327), th→en (299), ja→ko (163), lt→en (68), and zh→de (150). Across these seven directions, all 1,759 known incorrect translations were blocked from direct acceptance.

5.4 Automatic Post-Editing Analysis

Automatic Post-Editing was a major source of workflow recovery. Across the benchmark, DAAIT corrected 4,058 known incorrect segments, equivalent to 67.6% of the full test population. The table below highlights the eight language pairs with the highest APE recovery rates.

Highest APE Achieved					
Pair	**Segments**	**Fixed by APE**	**APE rate**	**Need HPE**	**HPE rate**
lt→en	68	57	83,8%	11	16,2%
en→ru	472	376	79,7%	95	20,1%
zh→en	1.209	883	73,0%	297	24,6%
en→es	725	514	70,9%	206	28,4%
ar→en	361	255	70,6%	106	29,4%
vi→en	391	276	70,6%	115	29,4%
ja→ko	163	114	69,9%	49	30,1%
ja→en	912	634	69,5%	262	28,7%

The strongest recovery result was observed for lt→en, where DAAIT repaired 57 of 68 incorrect segments, leaving only 16.2% for Human Post-Editing. en→ru also achieved a high APE rate of 79.7%. The zh→en result is operationally significant because 883 incorrect segments were repaired automatically, the largest absolute recovery volume among the listed pairs, despite that direction also contributing the greatest number of false acceptances. Overall, the results show that APE can substantially reduce the human-review queue, although recovery effectiveness remains language-pair dependent and should be monitored separately from MTQE rejection accuracy.

5.5 Cross-Script Translation Performance

One of the central challenges in multilingual Machine Translation Quality Estimation is evaluating translations between languages that use fundamentally different writing systems. Unlike related Latin-script languages, cross-script directions provide little or no surface lexical overlap between the source and target. Reliable evaluation must therefore depend primarily on semantic correspondence, adequacy, and error detection rather than orthographic similarity.

Despite the absence of shared writing systems, MTQE achieved correct-rejection rates between 97.6% and 100% for eight of the nine evaluated cross-script language pairs. False acceptance remained below 2.5% for every cross-script direction except Spanish→Japanese, and four cross-script pairs recorded zero false acceptances.

The evaluated directions span Latin, Han, Japanese Kanji and Kana, Hangul, Cyrillic, Arabic, and Thai writing systems. The consistently high rejection accuracy across these combinations indicates that the model can assess translation adequacy where source and target forms provide minimal surface-level cues. Within the limits of this benchmark, the results are consistent with MTQE relying on semantic alignment and translation-error evidence rather than simple lexical or orthographic resemblance.

5.6 Runtime Remains Stable Across Languages

Processing time was tightly clustered for most of the benchmark. Eleven of the 16 language pairs, representing 5,481 of the 6,006 segments, or 91.3% of the evaluated volumen completed in the 0.36–0.47 second-per-segment range. This group includes the largest datasets: zh→en processed 1,209 segments at 0.36 seconds each, en→es processed 725 at 0.37 seconds, and ja→en processed 912 at 0.47 seconds.

The remaining five pairs had higher averages: ja→ko at 0.65 seconds, es→ja at 0.97 seconds, ja→es at 1.15 seconds, zh→de at 1.19 seconds, and lt→en at 1.47 seconds. Four of these five datasets contained 150 segments or fewer, and the highest observed latency occurred on the smallest dataset, lt→en, with 68 segments. This pattern suggests that fixed request, file-processing, or setup

overhead may have a greater effect on averages for small samples; the benchmark does not isolate those components, so the result should not be interpreted as proof that language has no runtime effect.

Across all 6,006 segments, aggregate runtime was 2,944 seconds, equivalent to a weighted mean of approximately 0.49 seconds per segment. The concentration of more than 90% of benchmark volume below half a second indicates broadly stable throughput across diverse languages and scripts and supports multilingual production deployment, subject to validation under the intended concurrency, infrastructure, file size, and service-load conditions.

5.7 Directional Robustness Across Bidirectional Language Pairs

To examine whether translation direction influences MTQE performance, the benchmark included two random language combinations in both directions: German↔Chinese and Spanish↔Japanese. These combinations involve substantial differences in morphology, syntax, and writing system, providing a demanding comparison of whether the framework maintains reliable decisions when the source and target roles are reversed.

Bidirectional language-pair performance				
Language Pair	**Accuracy**	**Average Score**	**APE Recovery**	**HPE Rate**
German → Chinese	98.67%	78.30	61.3%	37.3%
Chinese → German	100.00%	53.99	66.0%	34.0%
Spanish → Japanese	93.16%	76.84	54.7%	38.5%
Japanese → Spanish	98.51%	76.05	58.2%	40.3%

Decision quality remained high in both directions, although the size of the directional effect differed by pair. German→Chinese achieved 98.67% correct rejection and Chinese→German achieved 100.00%. Spanish→Japanese recorded 93.16%, compared with 98.51% for Japanese→Spanish. Thus, reversing direction did not undermine the overall ability to reject most known incorrect translations, but the Spanish→Japanese result shows that direction can still materially affect residual false-acceptance risk.

Downstream recovery was more consistent. APE recovered 61.3% and 66.0% for the German–Chinese directions and 54.7% and 58.2% for the Spanish–Japanese directions. HPE rates ranged from 34.0% to 40.3% across all four tests. This comparatively narrow range suggests that, once an incorrect segment was routed for intervention, the automated recovery process behaved similarly in both directions.

These findings provide evidence of directional robustness, but they should be interpreted cautiously because the directional datasets are not equally sized: German→Chinese contains 75 segments versus 150 for Chinese→German, while Spanish→Japanese contains 117 versus 67 for Japanese→Spanish. Within those limits, MTQE maintained reliable quality estimation across both bidirectional scenarios despite major linguistic and script differences, while pair-specific results confirm that direction should remain a monitored deployment dimension.

5.8 Operational Impact Analysis

The benchmark presented in this study evaluates the technical performance of MTQE under controlled multilingual conditions. To illustrate the practical implications of these results, we model three representative translation workflows using an illustrative machine translation system that produces acceptable translations for 90% of translated segments. The remaining 10% contain translation errors requiring additional processing. This assumption is intended solely to demonstrate the impact of MTQE on an existing production workflow and does not represent the performance of any specific machine translation system.

The analysis applies the measured benchmark performance obtained in this study:

- Correct rejection accuracy: 98.9%
- False acceptance rate: 1.1%
- Automatic Post-Editing recovery rate: 67.6%

For simplicity, the calculations are based on 100 translated segments.

Scenario 1 – Machine Translation Only

Without quality estimation, all translations produced by the MT system are delivered directly to the downstream workflow. Assuming an MT quality of 90%, approximately 90 translations are immediately publishable, while 10 translations contain quality issues. Since no automated quality assessment is performed, all incorrect translations remain undetected unless every segment undergoes human review. Consequently, achieving production-quality output requires 100% human revision of the translated content.

Scenario 2 – Machine Translation with MTQE

Introducing MTQE fundamentally changes the workflow by automatically filtering translations before publication.

Of the 10 incorrect translations, the benchmark indicates that 98.9% are correctly identified and rejected. Therefore,

$$10 \times 0.989 = 9.89$$

incorrect translations are intercepted before publication, while only

$$10 \times 0.011 = 0.11$$

incorrect translations pass the quality gate.

The total number of translations accepted directly becomes

$$90 + 0.11 = 90.11$$

Of these, 90 translations are actually correct, resulting in an acceptance precision of

$$\frac{90}{90.11} \times 100 = 99.88\%.$$

In other words, while the original MT system produces translations that are correct approximately 90% of the time, the translations approved by MTQE are correct almost 99.9% of the time. Rather than publishing every translation, MTQE isolates only the potentially problematic 9.89% of the workload for additional processing, dramatically reducing publication risk.

Scenario 3 – Machine Translation with MTQE and Automatic Post-Editing

The rejected translations are subsequently processed using Automatic Post-Editing.

Based on the benchmark, 67.6% of rejected translations are successfully repaired.

Applying this recovery rate,

$$9.89 \times 0.676 = 6.69$$

translations are automatically corrected, leaving

$$9.89 - 6.69 = 3.20$$

translations requiring Human Post-Editing.

The estimated number of publishable translations therefore increases from the original 90 translations to

$$90 + 6.69 = 96.69,$$

corresponding to an estimated publishable quality of 96.69%.

Perhaps more importantly, the amount of content requiring professional review decreases dramatically. Instead of reviewing all 100 translated segments, linguists are required to review only approximately 3.2 segments, representing a reduction in manual review workload of

$$\left(1 - \frac{3.20}{100}\right) \times 100 = 96.8\%.$$

This demonstrates that MTQE combined with Automatic Post-Editing enables human expertise to be concentrated on the relatively small proportion of translations that cannot be resolved automatically.

Metric	MT Only	MT + MTQE	MT + MTQE + APE
Publishable quality	90.0%	90.0% + HPE fixed	96.69% + HPE fixed
Incorrect translations reaching publication	10.0%	0.11%	0.18%
Human review workload	100%	9.89%	3.20%

These calculations illustrate how MTQE transforms the localization workflow. Instead of relying on exhaustive manual review, the system first prevents almost all erroneous translations from reaching publication, then automatically repairs the majority of rejected translations, leaving only a small fraction for human linguists. The result is a workflow that simultaneously improves quality assurance and significantly reduces manual effort.

This operational analysis is based on a simplified deployment model and assumes that the performance of MTQE is independent of the quality of the underlying machine translation system. In practice, this assumption may not always hold. Translation errors that are particularly challenging for an MT system may also be more difficult for MTQE to detect, introducing correlation between MT failures and MTQE decisions. Similarly, the effectiveness of Automatic Post-Editing depends on the characteristics of the errors produced by the upstream translation engine. Consequently, the projected improvements presented here should be interpreted as an analytical estimate derived from the measured benchmark performance rather than guaranteed operational outcomes for every MT system. Future work will investigate the interaction between specific MT engines and MTQE to quantify these dependencies under real-world deployment conditions.

6. Limitations

Limited language coverage. Although ACES spans 146 language pairs, this study evaluated only 16 because many of the remaining directions contain too few examples for a meaningful assessment of MTQE robustness. Specifically, 99 of the 146 ACES language pairs have fewer than 100 test segments. Regional varieties, dialects, and additional bidirectional combinations also remain untested, so the aggregate result should not be generalized to unsupported directions without further validation.

Uneven sample sizes and English-centric representation. The number of segments varies substantially by pair, from 67 for ja→es and 68 for lt→en to 1,209 for zh→en. Consequently, an isolated accepted segment produces a visibly larger rate change in the smallest sets, and zero false acceptances in a small set provide less evidence than the same result over hundreds or thousands of examples.

Incorrect-only evaluation design. The benchmark intentionally focuses on known incorrect translations because the primary objective is to measure whether MTQE detects unsafe output. This design supports direct calculation of false acceptance and correct rejection, but it does not measure how often valid translations are unnecessarily rejected. As a result, the study cannot report false rejection, precision on acceptable output, overall accuracy across a realistic mix of good and bad translations.

Pipeline-level rather than component-only results. The reported recovery rate reflects the combined behavior of MTQE routing, DAAIT correction, and MTQE re-evaluation. It should not be interpreted as the standalone performance of either MTQE or DAAIT. Because translation memories and glossaries are not applicable for this task, the benchmark also does not quantify the potential benefits or risks of customer-specific terminology and retrieval assets.

These limitations do not negate the benchmark findings, but they define their scope. Before broad deployment, the framework should be validated on larger and more balanced language-pair samples, include both acceptable and incorrect translations, report confidence intervals and phenomenon-level results, and be tested on representative enterprise domains under production load.

7. Conclusion

This paper presented a production-oriented benchmark of Pangeanic MTQE v2 using 6,006 known incorrect translations drawn from the ACES challenge set across 16 language pairs. Rather than evaluating MTQE only as a continuous scoring metric, the study assessed its operational role: determining whether a segment could proceed, should be routed to DAAIT for Automatic Post-Editing, or required Human Post-Editing.

The results indicate strong error-detection performance under these controlled conditions. MTQE correctly rejected 98.9% of the known incorrect segments, with a 1.1% false acceptance rate. Seven language pairs recorded no false acceptances, and high rejection performance was maintained across diverse scripts and translation directions. The remaining errors were concentrated in a small number of language pairs, reinforcing the need for pair-specific monitoring rather than reliance on aggregate performance alone.

The benchmark also demonstrates the value of combining quality estimation with automated recovery. DAAIT corrected 4,058 segments, or 67.6% of the full incorrect-segment population, while 31.4% required Human Post-Editing. This layered workflow turns MTQE into more than a classifier: it becomes a controlled orchestration layer that detects risk, directs recoverable errors to correction, re-evaluates the result, and reserves human expertise for unresolved cases.

Runtime results further support operational use. More than 90% of the evaluated volume was processed within 0.36–0.47 seconds per segment, and the full benchmark completed at a weighted mean of approximately 0.49 seconds per segment. Together with reference-free scoring, explainable error descriptions, and API-based routing, this performance supports integration into high-throughput multilingual localization pipelines.

The findings should be interpreted within the scope of the study. The benchmark evaluates selected ACES incorrect translations rather than a production distribution containing both acceptable and unacceptable content. Language-pair coverage and sample sizes are uneven, and the results do not measure false rejection of valid translations or deployment-specific cost savings. Broader validation should therefore include balanced good and bad translations,

additional non-English directions, phenomenon-level analysis, representative enterprise domains, and production-load testing.

Within these limits, the benchmark shows that modern MTQE has evolved beyond quality scoring into practical decision support for trustworthy machine translation deployment. By acting as a multilingual quality gate, blocking known errors, enabling automated correction, and escalating only the remainder - Pangeanic MTQE v2 provides a credible foundation for risk-based, explainable, and scalable translation quality assurance.

8. References

Amrhein, C., Moghe, N., & Guillou, L. (2022). *ACES: Translation Accuracy Challenge Sets for Evaluating Machine Translation Metrics*. In Proceedings of the Seventh Conference on Machine Translation (WMT) (pp. 479–513). Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.44

Fomicheva, M., Specia, L., Blain, F., Chaudhary, V., & Guzmán, F. (2022). *MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset*. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022) (pp. 4963–4974). European Language Resources Association.

Lommel, A., Burchardt, A., & Uszkoreit, H. (2014). *Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics*. Tradumàtica, 12, 455–463.

Moghe, N., Amrhein, C., & Guillou, L. (2024). *ACES: Translation Accuracy Challenge Sets for Automatic Evaluation Metrics*. arXiv preprint arXiv:2401.16313. https://arxiv.org/abs/2401.16313

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2022). *COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task*. In Proceedings of the Seventh Conference on Machine Translation (WMT). Association for Computational Linguistics.

Specia, L., Blain, F., Fomicheva, M., Fonseca, E., Chaudhary, V., Guzmán, F., & Martins, A. F. T. (2020). *Findings of the WMT 2020 Shared Task on Quality Estimation*. In Proceedings of the Fifth Conference on Machine Translation (WMT) (pp. 743–764). Association for Computational Linguistics.

Chatterjee, R., Negri, M., Turchi, M., Federico, M., & others. (2017). *Guiding Neural Machine Translation Decoding with External Knowledge*. In Proceedings of the Second Conference on Machine Translation (WMT).

Pangeanic. (2026). *MTQE API Documentation*. Pangeanic Documentation. https://docs.pangeanic.com/mtqe-api/

Pangeanic. (2026). *Machine Translation Quality Estimation*. Pangeanic. https://pangeanic.com/machine-translation-quality-estimation-mtqe

9. Appendix

Due to the size of the benchmark (6,006 evaluated segments across 16 language pairs), only aggregate statistics are presented in this whitepaper. Complete per-segment results, MTQE outputs, and post-editing analyses are available in the accompanying SharePoint repository:

Repository: MTQE_ACES_Results