Association for Machine Translation in the Americas (AMTA)
Quality Estimation Users Working Group

Principles and Recipe

for Evaluating Quality Estimation (QE) Systems

August, 2026

Copyright, License, and Disclaimer

Executive Summary

Quality estimation (QE) systems are increasingly used to assess translation quality and support decisions about review, routing, and data selection without relying on reference translations. This document, developed by the AMTA Quality Estimation Users Working Group, provides practical **Principles and a step-by-step Recipe for evaluating QE systems from the user perspective**.

The central principle is that **QE performance is use-case and context specific**. Performance may vary by language pair, domain, content type, translation source, error distribution, and over time. Users should therefore evaluate QE systems under conditions representative of their actual workflows and avoid applying results or thresholds broadly without validation. Different QE use cases—including MT system comparison, segment-level operational routing, and automated error annotation—also require different evaluation approaches and performance indicators.

The Recipe turns these principles into a practical six-step process: **collect representative data and a gold standard; prepare and quality-check the data; select and run QE systems; analyze results; interpret findings and limitations; and decide on next steps**. It focuses primarily on segment-level QE for operational routing and has been tested in more than 10 projects involving over 10 commercial and open-source QE systems, including trained QE models and general-purpose LLMs prompted to perform QE, across a wide range of content types, language pairs, and risk-tolerance levels.

For operational decisions, the Recipe emphasizes analysis of **true positives, false positives, true negatives, and false negatives** to understand performance across decision thresholds and quantify trade-offs between review cost and quality risk.

The framework does **not** prescribe a universal threshold, metric, or acceptable level of QE performance. Instead, it helps users generate evidence for decisions based on their own data, quality requirements, risk tolerance, workflows, and business constraints.

Acknowledgements

The AMTA Quality Estimation Users Working Group gratefully acknowledges its members for their time, expertise, and contributions throughout the development of this work. Their insights, discussions, testing, feedback, and practical perspectives helped shape and refine both the principles and the evaluation approaches presented here.

Working Group members:

- Carola F Berger, CFB Scientific Translations LLC
- Marianna Buchicchio, Independent Researcher
- Matthew Carulli, Korn Ferry
- Evelyn Garland, Acta-Transphere
- Marcin Kotwicki, European Commission DG Translation
- Alon Lavie, Carnegie Mellon University
- Arle Lommel, MQM Council
- Robbie Numbers
- Mirko Plit, International Monetary Fund
- Ammon Shurtz, Brigham Young University
- Ingemar Strandvik, MQM Council
- Angelika Vaasa, Translation Service of the European Parliament
- Franziska Willnow, Amazon Web Services (AWS)

The Working Group also thanks the Association for Machine Translation in the Americas (AMTA) for hosting the group and for providing the resources, infrastructure, and forum that made this collaborative effort possible.

Special thanks are extended to Ammon Shurtz and Brigham Young University (BYU) for supporting Working Group members' access to open-source models used in testing and experimentation.

Table of Contents

AMTA Quality Estimation Users Working Group

The goal of the working group was to establish, recommend, and, at a later stage, help implement an effective and shared framework and methodology for evaluating Automated Translation QE systems.

This initiative was grounded in alignment with current technological realities. QE technology is evolving from being primarily based on trained neural systems toward GenAI-based reasoning models, and the proposed evaluation principles are intended to remain relevant across this transition.

The work of the group was also explicitly and intentionally use-case driven. Rather than evaluating QE systems in the abstract, we focused on how these technologies are actually used and tested in enterprise settings today. This included identifying key QE use cases, understanding how they differ, and clarifying how requirements for evaluating QE systems change depending on context, risk profile of the translated content, and operational constraints.

Direct execution of evaluations of current commercial QE systems was beyond the scope of the working group.

Principles

Introduction

The Principles provide high-level guidance on what to do (and what not to do) when evaluating quality estimation (QE) systems. They are intended to support QE system users to make informed decisions when evaluating QE system performance, and are not specifically designed to support QE system development or improvement.

The Principles are followed in this document by a step-by-step Recipe, also developed by the AMTA Quality Estimation Users Working Group, that provides practical guidance for their implementation.

Use cases

Quality Estimation (QE) evaluation methodology should be explicitly aligned with the intended operational use case. Below are three distinct QE use cases, each requiring different evaluation logic and performance indicators.

QE Use Case 1 (UC1): Contrastive quality evaluation of machine translation (MT) systems

Objective

Compare the relative performance of multiple MT systems on a benchmark dataset without relying exclusively on reference-based metrics.

Operational scenario

- Offline evaluation of MT systems (outside of the real-time production environment)
  - Conducted on a curated benchmarking dataset
  - Dataset typically controlled for aspects such as language pair (including direction), domain, and content type
- Online evaluation of MT systems (in a real-time production environment)
  - Once the reliability of a QE system is established offline, the QE may be used in a real-time production environment to select the best MT system

Decision formulation

- MT system-level comparison
- Produces aggregate conclusions about relative MT system performance

Gold standard

- Human MQM-based annotation; or
- Less ideal but may still be useful: reference human translation

Performance indicators

- Correlation with human scores, rating, ranking, or reference translation

Related WMT work

- One of the primary WMT Shared Task use cases
- Well-established methodology, see Section 4.2 of the WMT25 Shared Task Findings Paper

QE Use Case 2 (UC2): Segment-level quality scoring for operational routing

Objective

Identify segments that meet a defined quality threshold and may not require further review or may be used as high-quality segments for model training, or fail a defined quality threshold and may require further review or may be unsuitable to be used for model training

Operational scenario

- Real-time or near-real-time scoring at production time
- Used for workflow decisions (block, route, highlight, or prioritize for editing)
- Sensitive to content criticality and risk tolerance
- Variant: post-production (i.e. after human post-editing) translation quality evaluation

Decision formulation

- Typically binary or multi-class classification of segments
- Typically based on QE system score + selected threshold

Gold standard

- Human MQM-based annotation at the *segment level;*
- Variant: other human annotation that reflect translation errors at the *segment level* that matter to the user; or
- Less ideal but may still be useful: human edits, on the assumption that the presence of edits indicates translation errors, and the absence of edits indicates error-free segments

Performance indicators

- Correlation with human scores or rating
- Ability to separate error-containing translation segments from error-free translation segments
- Consistency of the QE system in detecting the same errors in the same dataset and across different datasets

Related WMT work

- Subtask 3 under WMT2026 Shared Task: Automated Translation Quality Evaluation Systems (new subtask started in 2026)

QE Use Case 3 (UC3): Fully automated error annotation for linguistic quality assessment (Auto-LQA)

Objective

Perform complete or partial automated error annotation and/or correction. Specific tasks include any combination of the following: identifying error span (location), error type, and/or error severity; and suggesting corrections of errors.

Operational scenario

- Real-time or near-real-time qualitative output at LQA time
- Used for augmentation of human LQA, pre-annotation to support human-verified LQA, or automated substitution for human LQA

Decision formulation

- Typically binary or multi-class classification of individual errors
- Additional considerations for partial matches when the qualitative outputs span a range (e.g., error span) or multiple criteria (e.g., error type, severity, and correction suggestion)

Gold standard

- Human MQM-based error annotation at the error-span level; or
- Variant: other human annotation at the error-span level that reflects translation errors that matter to the user

Performance indicators

- Correlation with human-annotated error spans, types, and/or severity
- Consistency of the QE system in detecting the same errors in the same dataset and across different datasets

Relation to other use cases

- QE systems that perform accurate automated error annotation (UC3) can be easily adapted and used for UC1 and UC2
- However, QE systems that do not perform well for UC3 may still be useful for UC1 and UC2

Related WMT work

- Included in WMT Shared Tasks; methodology details in Perrella et al. (2026)

The following principles apply to all three use cases unless otherwise noted.

1. Generalization

QE performance cannot be assumed to generalize across different operational contexts, as system behavior varies widely by language pair, content type, domain, and other factors—and can change over time. The evaluation of QE systems should be explicitly and intentionally **use-case specific**, both with regard to the translation use case and the QE use case.

Key principles

- **Use a multi-aspect evaluation rubric**
  - The evaluation rubric should account for multiple aspects that are likely to affect QE system performance.
  - The total **number of aspects** and the **definition of each aspect** should be customizable based on user needs.
  - Example QE rubric aspects include:
    - Language pair (including direction)
    - Domain and content type (defined by the user to reflect meaningful content clusters)
    - Input translation characteristics, such as
      - Human translation, human post-edited machine translation, or machine translation (MT)
        - If MT, which MT system
      - Translation error rate
      - Translation error distribution
    - Additional aspects that are important to the user, such as the target audience and language-level compexity. may affect QE system performance
- **Establish aspect-specific thresholds and conclusions**
  - The evaluation methodology should enable conclusions such as:
    - *“QE system X performs better for language pairs A, B, and C, for content type T translated by MT engine Y, in domain Z”*

    - When score thresholds are applicable, a pass/fail threshold should be established and verified for each QE rubric aspect combination, based on the quality requirements and risk tolerance level of the translated content as well as the purpose of the quality estimation. [UC2]
- **Avoid making overly broad generalizations**
    - QE system performance will likely vary significantly across:
        - Language pairs (including direction)
        - Domains and content types
        - Input translation characteristics (e.g., error rate and distribution in the input translation)
    - Performance observed in one setting should not be assumed to hold in another without testing and validation.
    - When score thresholds are applicable, avoid applying the same general threshold to different QE rubric element combinations without validation. [UC2]
- **Be cautious of result aggregation**
    - Avoid reporting only aggregate averages across aspects (e.g., across multiple language pairs or multiple content types) as these often hide critically important differences and variance.
    - Instead, examine how the results vary across different aspects.
- **Account for performance drift over time**
    - Results from one time point cannot be assumed to generalize to future time points.
    - QE systems should be evaluated periodically, as performance can and will change.

Additional considerations

- **Note on sentence-level QE systems**
    - Because sentence-level QE systems lack access to document-level and external context, they can evaluate only a subset of the factors that determine translation quality.
    - In particular, they cannot reliably assess phenomena such as cohesion, coherence, adherence to style guides, or consistency with external knowledge, reference documents, or other portions of the text.
    - As a result, sentence-level QE should not be interpreted as a comprehensive measure of translation quality in applications where contextual accuracy is essential.
- **Prioritize**
    - When it is not feasible to test all possible combinations of the aspects in the rubric:
        - Select the most important items within each aspect (e.g., top language pairs, top domains and content types).
        - Test all combinations of these selected items.
-

2. Sample dataset

Effective sample dataset design for QE system evaluation requires careful choices about size, content, and structure, guided by the evaluation objective and available resources. The sample dataset should preserve meaningful context, reflect real production data where possible, and

balance continuity with diversity, while accounting for error distribution and translation specifications. Additional factors such as data cleaning, chunking strategy, and the use of challenge datasets can materially affect QE outcomes.

Key Principles

- **Decide on an appropriate sample size**
  - There is no “magic number”; appropriate size depends on:
    - Resource availability
    - Evaluation objective
      - For workflow decisions, datasets should reflect naturally occurring error distributions.
      - For assessing error discrimination, targeted or “challenge” datasets may be more informative.
    - Error distribution
      - Highly uneven distribution may require a larger dataset.
  - Empirical evidence:
    - Approximately 500-1,000 segments or 5,000–10,000 words per combination of evaluation rubric aspects is typically considered a good starting point.
- **Include meaningful data**
  - What is considered meaningful may be use-case specific.
  - Preserve context as much as possible
    - Do not mix segments from different documents.
    - Use entire documents or coherent sections where feasible.
  - For non-document content (e.g. software UI)
    - Consider clustering text and incorporating metadata or multimodal context.
  - Balance continuity and diversity
    - Consider the trade-off between using fewer documents with longer excerpts (better continuity) and using more documents with shorter sections (greater diversity).
  - Real production data is preferred where available.
  - When synthetic data is used:
    - Consider whether its error distribution mirrors real production data.
    - Be explicit about how and why synthetic data is introduced and potential consequences of using synthetic data.

Additional considerations

- **Dataset cleaning**
  - Open questions for future work include:
    - Whether to remove tags, duplicates, or near-duplicate segments.
    - How to account for repeated errors or segment-length effects if data is not cleaned.
    - Whether and how to apply length thresholds.
    - How cleaning decisions affect context integrity.
- **Data chunking**
  - How benchmark data is chunked can significantly affect QE results. QE systems do not necessarily perform equally well on sentence-long segments and paragraph-long segments.

- **Dataset difficulty**
  - Content that is more difficult for MT is not necessarily more difficult for QE.
  - A higher input translation error rate does not necessarily imply higher QE difficulty. Rather, testing by working group members suggests that an input translation with fewer and less severe errors can be more challenging for QE systems.
  - QE systems' performance challenges are multidimensional, including but not limited to:
    - detecting true translation errors (true positives),
    - detecting truly correct translations (true negatives),
    - minimizing misclassification of correct translations as translation errors (false positives),
    - minimizing misclassification of translation errors as correct translations (false negatives),
    - classifying error types correctly,
    - classifying error severity correctly, and
    - detecting errors correctly at the error-span level.

LLM-as-a-judge QE systems - special considerations [Primarily UC3, but also UC1 & UC2]

- **Context and granularity**
  - QE evaluation should distinguish between:
    - The granularity of scoring (e.g. segment-level, error-span-level), and
    - The context provided to the QE system.
  - When the granularity of scoring and the context provided to the QE system are not clearly distinguished, e.g. feeding a large chunk of text to an LLM-as-a-judge QE system without specifying a finer granularity for scoring, LLM QE performance can deteriorate.
  - LLM-as-a-judge QE systems can benefit from broader contextual windows (e.g., full documents) beyond what has traditionally been sent to encoder-decoder NMT models, even when producing segment-level scores. But keep in mind that LLMs often fail at the limits of their context windows.
- **People-pleaser effect**
  - Some LLM-as-a-judge QE systems exhibit a tendency to generate false positives on error-free content when asked to identify translation errors (the "people-pleaser" effect).
- **Lack of consistency**
  - Unlike deterministic metrics such as COMET, LLM outputs can be inconsistent: the exact same prompt may yield different results, sometimes with substantial divergence.
  - This inconsistency can be mitigated by setting the temperature parameter of the LLM to 0. However, even at temperature 0, strict consistency is not guaranteed. Additionally, many recent reasoning models do not allow the temperature to be set to 0.
  - Therefore, LLM output consistency should be verified and assessed against an acceptable range.

3. Gold Standard

A gold standard is the best available benchmark against which QE system outputs are evaluated, and its design has a direct impact on the validity of evaluation results. The appropriate type of gold standard depends on the specific use case and the evaluation goals. Gold standards are inherently imperfect and must be carefully designed, documented, and quality-controlled.

Key Principles

- **Select the right type of gold standard**
  - Choose a gold standard that aligns with the use case and evaluation goals.
  - Common options include
    - Multidimensional Quality Metrics (MQM) annotation
    - Error Span Annotation (ESA)
    - Translation Edit Rate (TER) against a reference translation (often the easiest to obtain)
    - Other labels that indicate translation errors that matter to the user
  - Decide whether the gold standard is
    - Fully human-produced, or
    - Machine-derived and confirmed by humans
  - Balance what is theoretically ideal with what is practically achievable.
- **Define what counts as a translation error**
  - Specifications are needed to define what constitutes a translation error based on the purpose and risk profile of the translation
  - Define error categories and severity levels
    - In addition to MQM, alternative error categorizations can support QE evaluation, such as:
      - Binary and non-binary errors (Pym 1992; Strandvik 2025)
      - American Translators Association certification exam error categories and error point values
  - Consider extrasegmental errors (i.e. issues that are errors because of something outside the segment, such as style guides, termbases, and reference documents; typical extrasegmental errors include lack of cohesion and inconsistency), and whether the QE system is capable of detecting extrasegmental errors.
- **Specify how translation error status is determined**
  - Specify the cutoff that separates what is acceptable (“pass”) and what is not acceptable (“fail”), if applicable
- **Acknowledge and manage imperfections**
  - Gold standards are almost never perfect.
  - Annotators may miss errors—including major ones—or incorrectly label non-errors as errors.
  - Measures should be taken to assess and improve gold standard quality, including
    - Completeness, i.e. whether the gold standard covers all aspects important to the user
    - Accuracy
    - Consistency, across different annotators and over time for the same annotator
- **Ensure appropriate human annotator qualifications**
  - Annotations should be performed by individuals whose skill level matches or exceeds that of those who normally sign off on the translation.
    - For example, if professional linguists are normally required to sign off on the translation, students or crowd workers should not be used as annotators.

- **Provide clear guidance**
  - Without shared specifications and guidance, annotation consistency is unlikely.
  - Human annotators should have access to
    - Project specifications for the translated material.
      - If such specifications do not exist, they should be inferred and documented.
    - Evaluation guidelines and decision trees (e.g., MQM decision trees).
  - Provide training to human annotators as needed
    - For example, when new annotators are recruited, or when identified borderline cases need to be clarified.

Additional considerations

- **Context-aware gold standard**
  - Where feasible, create a gold standard that takes the context of the translated content into consideration, even if the original MT process is strictly segment-based.
  - Providing broader context will better reflect real-world usage and may reveal patterns not visible in isolated segment evaluation.
- **Trade-off between error classification effort and analytical depth**
  - While classifying errors into multiple types and severity levels can yield richer insights, a simpler classification approach—error vs. no error—may be sufficient for certain use cases.
  - Even in this simplified approach, errors should be clearly defined, and the criteria for determining error status should be specified.
- **Ongoing validation**
  - Periodically review and validate the gold standard, especially for long-running evaluations or repeated testing over time.
  - Treat the gold standard as a controlled but evolving benchmark, not an unquestionable ground truth.
- **Public benchmark caveats**
  - Reuse of public benchmarks can be problematic because they are often exposed and may have been incorporated into the training data of QE models.
  - Creating new benchmarks for each evaluation reduces this risk but complicates longitudinal comparison over time.
  - An alternative is curating a private benchmark, which can be reused but still need to be updated over time as MT quality improves and the user's content evolves.

4. Performance Indicators

Robust QE system evaluation depends on clear definitions, explicit decision rules, and alignment with the intended use case. Each evaluation project should have a strategy to define what constitutes a translation error, how error status is determined, and which QE capabilities are being assessed, then select performance indicators accordingly.

Key Principles

- **Distinguish QE capabilities** [Primarily UC3, but also UC1 & UC2]
  - Treat the following as separate capabilities and evaluate them independently:
    - Error detection

        - Error annotation (error type and severity)
        - Error correction suggestion
- **Be explicit about evaluation focus**
    - Specify the QE use case and its objective
    - To evaluate the performance of a QE system for different use cases and objectives, different performance indicators are needed.
- **Select appropriate performance indicators**
    - Choose performance indicators that align with the use case.
    - There is no single "best" performance indicator for all use cases.
    - Common performance indicators include but are not limited to correlation coefficients, recall and precision, and ROC-AUC.
    - The accompanying Recipe provides more details on performance indicators for various use cases.

Additional considerations

- **Assess usefulness, not just correctness**
    - QE outputs can also be evaluated based on whether they are useful to users, regardless of the output format (score, detection, error span, type, severity, correction suggestion, etc.).
    - Consider ergonomics for human-oriented use cases, for example
        - A large number of false negatives[1] can lead to a perception that the QE system does well, but it is actually missing errors; in contrast, a large number of false positives[2] can lead to a perception that the QE system underperforms or is not useful, even if the system catches more errors.
- **Understand qualitative QE outputs** [primarily UC3, but also UC1 & UC2]
    - Many recent QE systems can produce qualitative outputs—such as error span, type, severity, and correction suggestion—rather than a single quality or confidence score.
    - In these cases, an implicit probability threshold is still associated with each qualitative output; that is, each output may be correct or incorrect and therefore requires validation.
- **When the confusion matrix is used, be explicit about how elements of the confusion matrix** – true positives, true negatives, false positives, and false negatives – **are counted** [UC2 & UC3]
    - The element counts vary depending on
        - QE system output type
        - Strictness of matching between QE output and the gold standard
    - For score-based QE systems
        - Pass/fail QE score thresholds determine the counts for true positives and false negatives.
    - For QE systems that produce qualitative outputs spanning a range (e.g., error span) or multiple criteria (e.g., error type, severity, and correction suggestion)
        - True positives may be defined as
            - Full matches across all criteria;
            - Full matches across a subset of criteria; or
            - Partial matches across all or a subset of criteria.

[1] See the attached glossary for the definition of "false negatives."
[2] See the attached glossary for the definition of "false positives."

- Stricter matches generally reduce true positive counts, while more relaxed matches may increase ambiguity and lead to user frustration.

5. Future Readiness

Future-ready QE system evaluation should anticipate both technological evolution and increasing scale. As QE capabilities and translation workflows continue to evolve, evaluation frameworks should be designed to remain adaptable, scalable, and aligned with real-world production environments.

Key Principles

- **Plan for automation**
  - Automating parts of QE system evaluation enables testing across a broader range of combinations within the evaluation rubric.
  - When designing large-scale evaluations, identify which steps can be reliably automated.
  - QE system evaluation methodologies should remain flexible enough to accommodate new types of QE outputs and workflow integrations.
- **Account for technology evolution**
  - QE systems are evolving from segment-level scoring toward more granular outputs such as error span/type/severity detection, correction suggestion, and explanation.
  - In parallel, translation workflows are shifting from classic MTPE CAT-based models to GenAI-driven, agentic workflows that perform multiple quality assurance (QA) functions.
  - Periodic re-evaluation is needed
    - Conclusions drawn from current QE system evaluations may not fully hold as technologies evolve.
  - Consider how future QE systems will fit into existing or planned translation and evaluation workflows, rather than assuming static system behavior.

6. Risks

Using QE systems in production entails risks arising from inherent uncertainty in QE system outputs and from the need to generalize evaluation results beyond tested conditions. These risks—particularly the risk of false negatives (missed translation errors)—should be explicitly acknowledged and assessed.

Key principles

- **Understand limitations of QE systems**
  - QE systems estimate whether a translation error is present; these estimates can be incorrect.
  - QE system outputs should therefore not be treated as unchallengeable truth, even when the system reports high or complete confidence.
- **Acknowledge and assess generalization risk**
  - In practice, some degree of generalization is often unavoidable.
  - When evaluation results are generalized beyond the tested conditions, the potential impact on accuracy and risk should be assessed.
- **Beware of missed translation errors (false negatives)** [UC2 & UC3]

  - When false negatives pose a significant risk in QE-driven decision-making, they should be explicitly assessed as part of the evaluation process.
  - In production, periodically sample content classified as “not containing errors” or “no review needed” and manually inspect it to monitor false negatives.
- **Beware of non-errors mistakenly identified as errors (false positives)** [UC2 & UC3]
  - False positives, by nature, are unlikely to be missed; rather, the challenge is to avoid too many false positives.
  - Too many false positives incur excessive processing cost (time and/or money) and can lead to human users losing confidence in the QE system.
  - In production, monitor the percentage of false positives.

Recipe

Introduction

Automated translation quality estimation (QE) systems are increasingly used to assess translation quality without relying on a reference translation. These systems have been referred to by different names. Here, we refer to them as QE systems.

The Recipe focuses on Use Case 2 as described in the Principles, which is the most common use case among the Working Group members. It can also be adapted to partially cover Use Cases 1 & 3. It can support comparisons across different QE systems, help assess the risks and costs associated with using QE outputs in production, and guide decisions about how QE should be integrated into translation review workflows.

The Recipe has been tested by members in more than 10 test projects involving over 10 QE systems, including systems specifically trained for QE and general-purpose LLMs prompted to perform QE, spanning both commercial and open-source systems across a wide range of content types, language pairs, and risk-tolerance levels.

The guidance that follows distills lessons from this collective experience into a practical, adaptable framework that users can apply and refine in their own settings. It is not intended to be applied rigidly, serve as a one-size-fits-all protocol, or prescribe what constitutes acceptable QE performance. Instead, users are encouraged to adapt the steps, thresholds, metrics, and analysis methods to their own data, workflow, quality expectations, and business constraints.

Step 1: Collect linguistic data

1.1 Obtain representative sample text (per language pair)
- 500-1,000 source text segments;
- Target text of the 500-1,000 source text segments

1.2 Obtain a gold standard (definition below)

Commentary

Representative sample text

A representative sample is a dataset—typically a subset of the working dataset (i.e. data to be evaluated), or a separate but smaller dataset—that provides estimates of characteristics and results substantially similar to those that would be obtained from the working dataset.

In general, a representative sample should be sufficiently similar to the data it is intended to represent. This means that the sample and the working dataset should involve the same language pair, domain, and content type. In addition, the target text in both the sample and the working dataset should be produced by the same machine or humans with the same competence profile using the same translation methods.

Source text

For document-based content, the source segments should be continuous and sampled from one document or a few similar documents; for non-document-based content, the source segments should come from the same or similar contexts.

Target text

Target text segments are the translated segments whose quality you want to assess; they can be either machine translation (MT) or human translation (HT).

Special cases

If your data contains a significant amount of the following types of segments, keep them in the sample but exclude them from the segment count:

- Duplicate segments; or
- Minimal-risk segments, i.e. segments highly unlikely to contain errors; for example, segments that consist of single numbers

Tag-heavy content and fragmented content are more challenging for quality estimation.

Gold standard

A gold standard is the best available benchmark for your task. In reality, it is often imperfect but should be robust enough for the user to confidently rely on. Examples of gold standard include:

- Edits made to the target text segments in a human review process
    - No-edit segments = error-free translations
    - Edited segments = error-containing translations
- MQM-annotated data indicating the translation error types and severities that matter to the user
    - For example, the user may be interested in major errors only instead of all errors
- Other types of labeled data where the labels indicate translation quality issues that matter to the user
    - For example, customer complaints, visitor conversion rates

When the gold standard consists of numeric scores but the user needs to make a binary decision—for example, whether a translation is acceptable or unacceptable—the user can choose a threshold that converts the numeric scores into the two decision categories.

Note on MQM

MQM (Multidimensional Quality Metrics) provides a structured framework for translation quality evaluation. Rather than producing only an overall score, it identifies errors by type and severity, generating diagnostic information for quality management and root-cause analysis. Its shared error typology and scoring methodology enable evaluation results to be compared consistently across reviewers, projects, and language pairs, while still depending on expert

judgement to interpret linguistic context and ambiguity.

MQM also supports the development of QE systems, since most QE models learn from human judgments. By capturing not just acceptability but the nature and severity of errors, MQM produces data that enables more meaningful analysis of QE system performance. Even where full human annotation is too costly for routine use, it remains a key reference for validating automated metrics against genuine translation quality rather than superficial statistical correlations.

MQM is customizable. It is not a single fixed error typology that must always be used in full. It is a framework from which users select error types suited to their evaluation purpose and quality requirements. Different MQM profiles and scoring models serve different purposes, so scores across implementations are not always directly comparable.

As the MQM framework continues to evolve, the official MQM website remains the best source for current guidance.

Step 2: Prepare data

2.1 Create a Data Spreadsheet
- See example in Appendix 1

2.2 Clean the linguistic data (optional)
2.3 Perform quality checks (QC) on the linguistic data

Commentary

Create a data spreadsheet

We suggest presenting the data in a human friendly format when the original data format is not intended for human review, such as json. This will allow humans to review the data for patterns and insights and facilitate QC on the linguistic data.

Clean the linguistic data

How the linguistic data is cleaned should depend on the user's purpose and translation workflow. It may be desirable to remove tags and formatting in some but not all cases. Other data cleaning steps may be taken depending on the user's needs.

Perform QC on the linguistic data

Check for potential issues in the linguistic data, such as wrong target language, misalignment, and missing fields.

Step 3: Select and run QE system(s)

3.1 Select the QE system(s) you want to test
3.2 Run the QE system(s)

3.3 Export the QE output
- Fill in the QE system output columns in the Excel spreadsheet

3.4 Perform QC on the QE system output

Commentary

Run the QE system(s)

If the QE system takes translation specifications (style guide, glossary, etc.) as input, prepare and provide such specifications to the QE system.

Perform QC on the QE system output

Look for abnormalities, such as empty cells, unusual output numbers or texts, and misalignment.

Types of QE systems

Generally speaking, QE systems fall under the two broad categories below:
- Systems that are specifically trained to perform QE tasks
  - Commercial
  - Open-source
- General-purpose LLMs that are prompted to perform QE tasks

Step 4: Analyze results

4.1 Select an analytical method
- Classification-outcome-based methods
  - Evaluating thresholded binary decisions using the confusion matrix, i.e. true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs); also known as confusion-matrix-based methods
  - Best for scenarios where binary decisions are made, such as workflow triaging
  - **See Appendix 2 for details**
  - **See Appendix 3 for special cases seen in real production**
- Correlation-based methods
  - Evaluating how well QE scores track or agree with a gold standard
  - Often used for developing QE systems
  - Examples: Pearson, Spearman, and Kendall correlations, as well as their variants

4.2 Pre-process to obtain proper input for analysis

4.3 Create plots and calculate values of indicators

4.4 Evaluate statistical significance

Commentary

Focus on classification-outcome-based methods

The focus of this recipe on this type of method reflects the interest of members of the AMTA

Quality Estimation Users Working Group. Other methods can be highly useful as well. The userfulness of a method depends on the user's needs.

Correlation-based methods

These methods are well established. We refer users to the Conference on Machine Translation (WMT) papers for details about correlation-based methods. For example, Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems.

Pre-process to obtain proper input for analysis

For classification-outcome-based methods, this may include:

- Error determination
  - Classifying all segments as containing errors that matter vs. not containing errors that matter
- QE score normalization
  - Converting QE scores output by different QE systems to the same range
  - Unnecessary for some analytical methods, e.g. ROC analysis

Appendix 2

This appendix

- Guides the user in their selection of analytical methods based on what questions they want to answer
- Visualizes analytics with illustrative plots
- Explains the "how" and "why"

Step 5: Draw conclusions

5.1 Interpret findings of the analysis
- See Appendix 2 for examples

5.2 Identify limitations

Commentary

Generalizability

A common limitation of tests of QE systems is generalizability. The performance of a QE system often varies across languages, domains, content types, and user specifications. Findings from tests that follow this Recipe should not be over generalized.

Step 6: Decide on next steps

6.1 If the tested QE system does not meet the user's requirements
- Test more QE systems while maintaining budget for current translation practice

6.2 If the tested QE system meets the user's requirements
- Use the QE system and continue to monitor it

6.3 If the tested QE system shows potential of meeting the user's requirements
- Decide on next steps case-by-case

Commentary

What if none of the QE systems tested meet the user's requirements?

It is entirely possible that no QE systems meet the requirements of a particular user. In this case, this Recipe helps the user obtain and communicate evidence to show why no QE systems meet their requirements.

Next steps to consider when the tested QE system shows potential

- Run the QE system on the same dataset multiple times to determine repeatability of QE system performance
- Increase sample size
- Include more language pairs
- Include more content types
- Review the gold standard to confirm its reliability

Appendix 1
Data Spreadsheet

Commentary

- The suggested columns below are meant to be a starting point. As indicated, some columns are necessary for QE outputs or analysis. Other columns are suggested for data governance, traceability, and quality checks. The user should feel free to add or remove columns according to their needs.
- The suggested columns are not meant to be filled out all at once. Rather, they should be filled out in stages as the test progresses.
- Subsequent steps and automation should be taken into consideration when designing the data spreadsheet.

Column heading	Description	Necessary for QE outputs?	Necessary for data analysis?
origin	Where the linguistic data comes from		
collected_by	Linguistic data collector - name or ID		
prepared_by	Linguistic data preparer - name or ID		
doc_id	Document ID	✓	✓
seg_id	Segment ID	✓	✓
source_lang	Source language	✓	
target_lang	Target language	✓	
source_text	Source text	✓	
target_text	Target text	✓	
translation_type	Type of translation: machine or human		
edit_type	Post-edited machine translation or edited human translation		
skip	Indicate whether the segment should be skipped for QE system output or analysis (applicable in cases where the user decides to exclude segments based on criteria such as duplication, etc.)		
translator	Name or ID of the machine translation engine or human translator		
editor	Name or ID of the machine translation post-editor or human translation editor		
gold_standard	Record the gold standard here. For “gold standard” definition and description, please refer to the commentary under the recipe step of “collect linguistic data.”		

error_determination	This field indicates whether the target text contains any errors that matter to the user as compared with the benchmark. It is typically binary, such as “error” and “no error.” See also: commentary under the recipe step of “analyze results.”		✓
linguistic_data_signoff	The person responsible for linguistic data QC signs off here after confirming that the linguistic data in the previous columns passes QC and is ready to be used for subsequent steps.		
qe_output_[system 1 name]	Output from QE system 1. Some QE systems produce multiple outputs such as error spans, error types and severities, and scores; in this case, multiple columns may be created to record these outputs separately.		✓
[…]	Output from additional QE systems, one column for each system		✓
qe_output_signoff	The person responsible for QE output QC signs off here after confirming that the QE output is correct and ready to be used for subsequent steps.		

Appendix 2
Data Analysis Guide for Classification-outcome-based Methods

How to use this guide

Select from below the business question of interest to you. Each business question is followed by plots (visuals), one or more examples, and a brief explanation of the rationale. Some special cases seen in real production are described in Appendix 3. Terms used in this guide are explained in Appendix 4.

Commentary

There is always some uncertainty inherent to the QE system output. Even when the system is designed to output definitive labels, such as “error,” “no-error,” or specific error spans, types, or severities, there is a hidden, internal threshold that the system uses to produce its outputs.

In this guide, we focus on QE systems that output at least a numeric range of segment-level quality scores. If the QE output is a definitive label, some of the curves in the plots below would be a single point instead of a curve, and plots involving QE score thresholds would not be applicable.

We use the number of error-containing segments, errors at the segment-level, as a measure of the amount of translation errors. It is possible to measure and estimate translation errors at individual-error level, provided the QE system can reliably identify such individual errors.

We make the following global assumptions:

- all translation errors are caught and corrected in the review step;
- “positive” means the segment contains one or more translation errors, and “negative” means the segment is error-free;
- higher QE score means more “negative,” i.e. better translation; and
- the cost discussed below refers to the variable cost associated with reviewing the translation and does not include cost of ownership of QE systems

In scenarios where such assumptions do not hold, adjustments and allowances can be made to account for such deviations.

The charts and numbers below are for illustrative purposes only and shall not be generalized to different datasets.

All plots below and the underlying data tables can be created manually with a spreadsheet application such as Excel or automatically with a computer programming language such as Python, with the necessary columns in the Data Spreadsheet as shown in Appendix 1 as the input.

ILLUSTRATIVE EXAMPLES[3]

In the examples below, the score ranges of all QE systems are normalized to [0, 100], with 0 indicating no meaning preserved in the translation and 100 indicating perfect translation.

1. Comparing QE systems

❓*Which QE system performs best for my content and scenario?*

▶️Plot: ROC curves (Figure 1)

💡Example: In the figure on the right, the performance ranking is QE System 1 > QE System 2 > QE System 3.

The more "northwest" the ROC curve is, the better the QE system performs.

✅Why: For a given FPR, the higher the TPR, the better the QE system. System 1 has a higher TPR at almost every FPR, as compared to Systems 2 and 3.

For more information about using ROC curves to evaluate QE system performance, refer to Garland and Berger (2026).

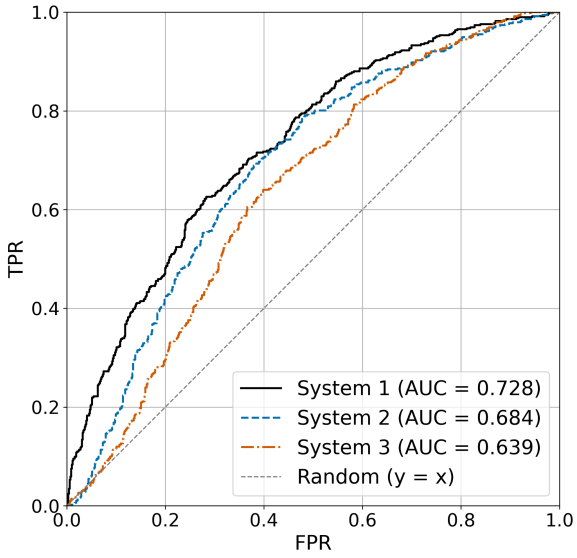
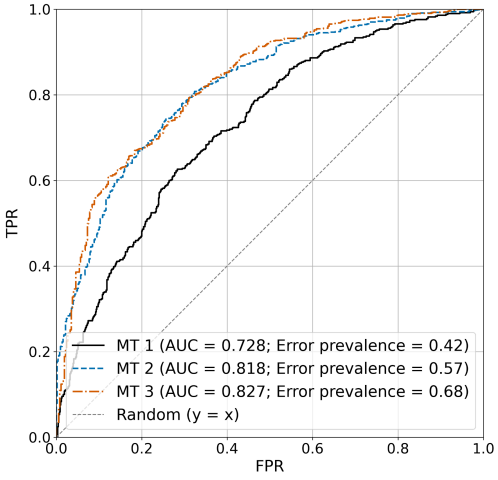


Figure 1

Additional Notes

A. QE system evaluation must be translation-specific

When the same QE system is applied to score different translations, a better ROC curve is often associated with a worse translation. For example, in the right figure, the QE System 1 from Figure 1 is used to score three different machine translations (MTs). MT 1 is higher quality compared to MT 2 and MT 3, as indicated by their error prevalence rates. However, the ROC curve associated with MT 1 is inferior to those associated with MT 2 and MT 3.



Therefore, QE system evaluation must be translation-specific. In other words, if a user uses MT 1 in production, they should evaluate

[3] The data underlying the examples is drawn from WMT23; see Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent (Freitag et al., WMT 2023). The authors gratefully acknowledge WMT for making extensive research data publicly available and for its contributions to advancing research in machine translation and quality estimation.

QE systems on that MT and not other MTs or human translations. When reporting ROC-AUC of a particular QE system, it is important to also describe the underlying translation, including its error prevalence rate.

2. Meeting a cost target and understanding the risk

❓*Which segments should be prioritized for review to meet a given budget? (Cost)*

▶️Plot: cost - QE score threshold (Figure 2)

💡Example: Suppose your budget is only enough to pay for 60% of all segments to be reviewed. Reading from Figure 2, the corresponding QE score threshold is 91. Send all segments whose QE scores are at or below 91 for review.

✅Why: A QE score threshold separates all segments into two subgroups: those that are more likely to contain errors, and those that are less likely to contain errors. By setting a QE score threshold, you control the relative size of the two subgroups. By reviewing the segments that are more likely to contain errors, you allocate review resources to the most needed segments.

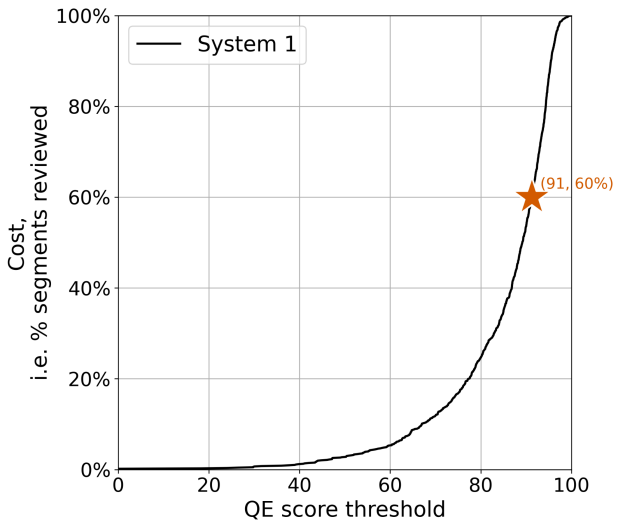


Figure 2

❓*How many error-containing segments will remain uncorrected? (Risk)*

▶️Plot: error prevalence - QE score threshold (Figure 3)

💡Example: When the QE score threshold is set at 91 to meet the cost target in the example above, error prevalence at the end of the review will be 0.09, i.e. 9 uncorrected error-containing segments per 100 segments.

✅Why: No QE system is perfect. A QE system can fail to detect errors in some of the segments. In this example, the QE score threshold is set at 91 to meet a cost target, and those segments above the threshold of 91 can still contain translation errors. Since they are not reviewed, these errors are not caught and corrected.

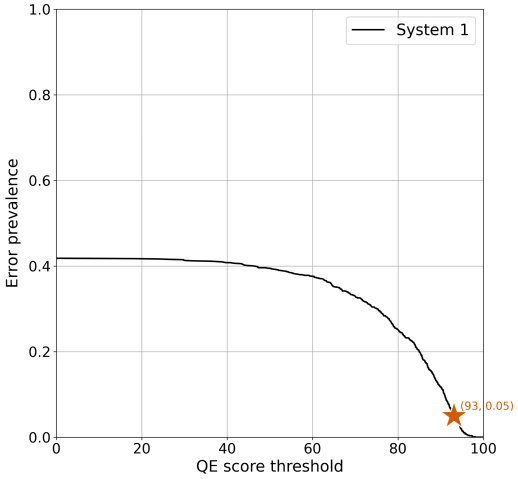


Figure 3

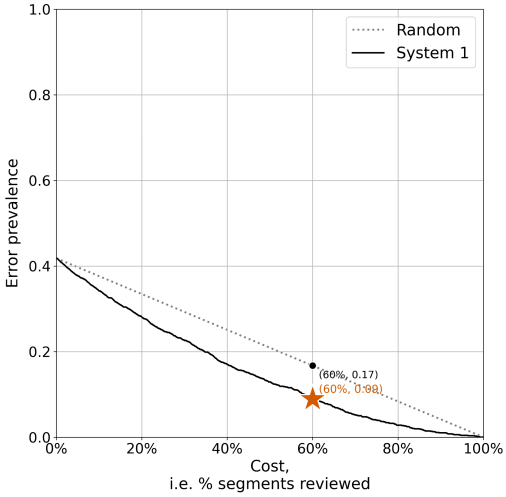


Figure 4

❓*What if I do not use a QE system?*

▶️Plot: error prevalence - % segments reviewed (Figure 4)

💡Example: You can choose segments to be reviewed randomly. As shown in Figure 4, reviewing 60% of the segments randomly lowers the error prevalence to 0.17 in this case, which is less efficient.

✅Why: A good QE system helps identify segments that are most likely to contain translation errors. You can achieve efficiency by focusing your resources on higher-risk segments.

3. Meeting a risk target and estimating the cost

❓*My risk tolerance is x error-containing segments per 100 segments. Which segments shall I send for review? (Risk)*

▶️Plot: error prevalence - QE score threshold (Figure 5)

💡Example: Suppose your error tolerance is 5 error-containing segments per 100 segments. Reading from Figure 5, the corresponding QE score threshold is 93. Send all segments with QE scores at or below 93 for review.

✅Why: The higher you set the QE score threshold, the more segments are sent for review, and thus more translation errors can be caught and corrected - keep in mind that no QE system

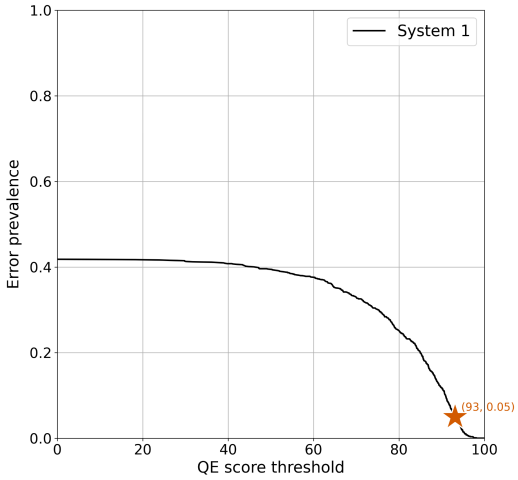


Figure 5

is perfect and even a 100-QE-score segment can contain translation errors. Once you know the behavior of a particular QE system on a particular type of content, such as the figure on the right shows, you can find a QE score threshold that corresponds to your risk tolerance.

❓*How much does it cost? (Cost)*

▶️ Plot: cost (i.e. % segments reviewed) - QE score threshold (Figure 6)

💡Example: Reading from the figure on the right, at a QE score threshold of 93, the cost is 71% of what it would cost to have all segments reviewed.

✅Why: See the explanation in the previous section under “Plot: cost - QE score threshold.”

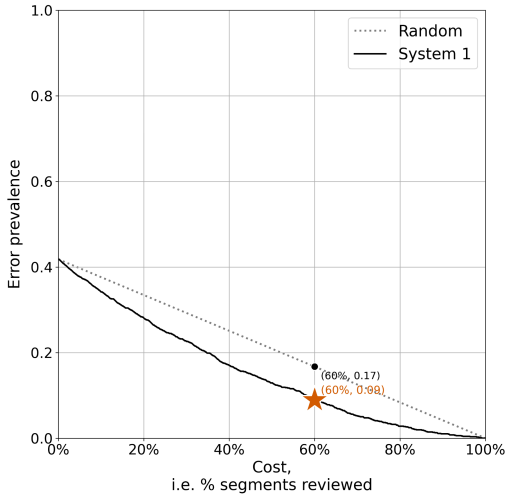


Figure 6

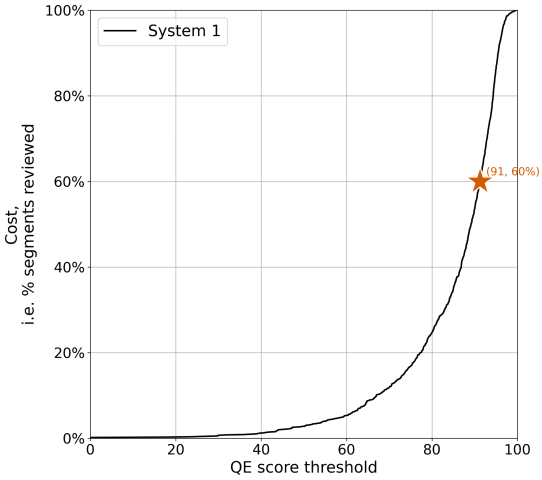


Figure 7

❓*What if I do not use a QE system?*

▶️Plot: error prevalence - % segments reviewed

💡Example: You can choose the segments to be reviewed randomly. With this approach, as shown in the figure on the right, to reduce error prevalence to 0.05 at the end of the review, 88% of segments need to be reviewed, which is less efficient.

✅Why: See the explanation in the previous section under “Plot: error prevalence - % segments reviewed.”

4. Balancing risk and cost

❓*My budget and risk tolerance are somewhat flexible. I have chosen a QE system, and I want to maximize its performance. What should I do?*

▶️Plot: ROC curve, and TPR - QE score threshold

💡Example 1 (orange): You have chosen QE System 1. You have determined that each segment containing translation errors that is missed (false negative, FN) costs you 3 times more than reviewing an error-free segment that is mistakenly flagged as containing error (false positive, FP). That is, 1 FN is equivalent to 3 FPs, and the FN-FP trade-off is 1:3.

Your MT system produces a translation that has 41% of the segments containing translation errors (i.e. among a total of 1,177 segments, 492 contain errors, and 685 are error-free; in other words, the P-N ratio is 492:685).

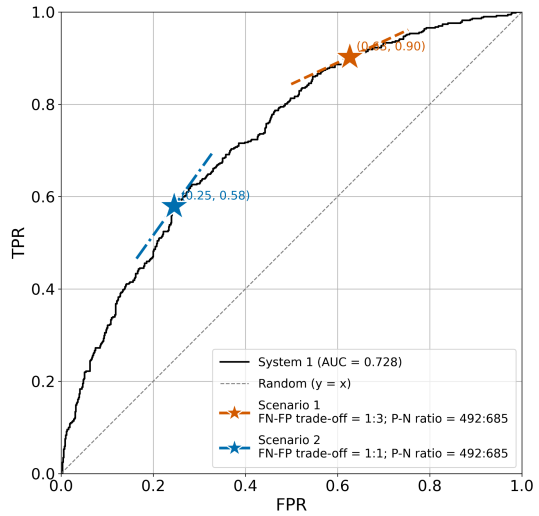


Figure 8

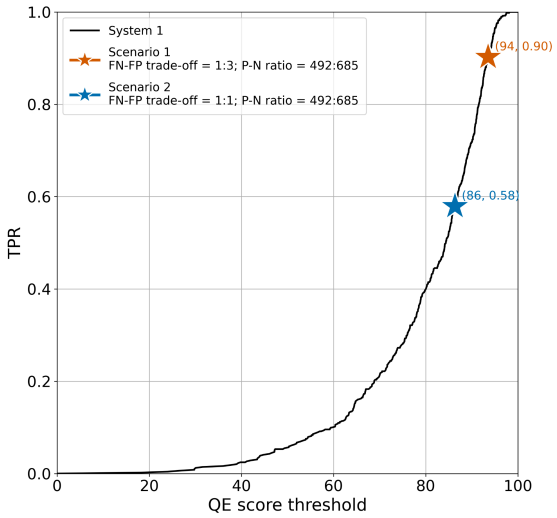


Figure 9

Create an ROC curve, and draw a line with a slope that equals the FN-FP trade-off divided by the P-N ratio. Find a point on the ROC curve where the line touches the curve tangentially and is as far northwest as possible. This point is QE System 1's optimal operating point for the given MT and your risk-cost trade-off.

Read the TPR value of the optimal operating point from the ROC curve, and find its corresponding QE score threshold from the TPR-QE score threshold curve. In this case, the QE score threshold is 94. That is, for your risk-cost trade-off and your chosen MT system, the QE system performs best when the QE score threshold is set at 94.

💡Example 2 (blue): Everything is the same as in Example 1, except that your FN-FP trade-off is 1:1. The blue star indicates the optimal operating point, and the optimal QE score threshold is 86.

✅Why: In general terms, for a given QE system, you cannot simultaneously reduce both false negatives (segments containing translation errors that are missed) and false positives (error-free segments that are mistakenly flagged as containing error). You can only trade one for the other. So identifying the optimal operating point involves figuring out what this trade-off is for you.

Why do you also need the P-N ratio? Simply put, the ROC curve shows the relationship between two *rates*, i.e. TPR and FPR, and you need the P-N ratio to convert the relationship between the rates to the relationship between the counts of false negatives and false positives.

For more details, refer to this paper.

5. Recall and precision

Recall and precision are also commonly measured indicators. Although they are not directly related to business outcomes such as risk or cost, they offer another perspective on the performance of QE systems: recall shows the percentage of all translation errors that are classified as errors by the QE system; and precision indicates the percentage of all segments classified as containing translation errors by the QE system that actually contain translation errors.

Recall and precision can be plotted against each other as a PR curve. Different QE systems exhibit different relationships between recall and precision. As illustrated in Figure 10, a strong QE system (System 1) is typically capable of achieving higher recall and precision simultaneously as compared to other QE systems (Systems 2 & 3). Note that in our example, as is frequently the case in general, the QE system with the best PR performance is also the one with the best ROC performance.

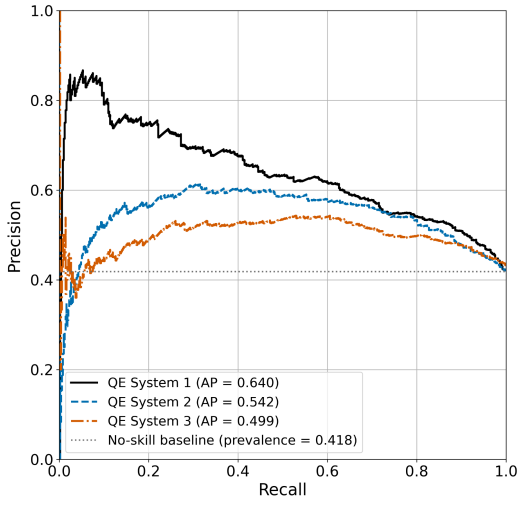


Figure 10

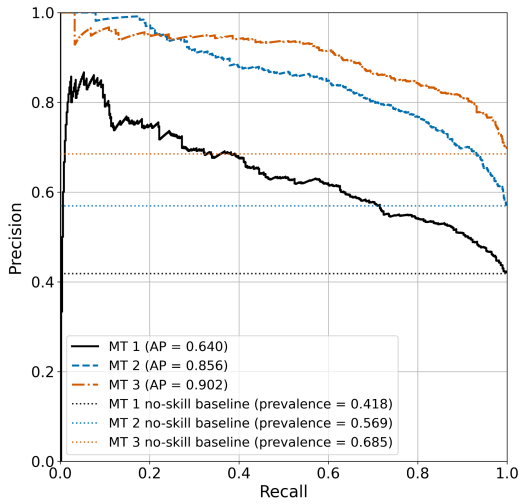


Figure 11

The user should keep in mind that the PR curve changes when the prevalence of translation errors changes. For example, in Figure 11, the same QE System 1 exhibits very different PR curves when used to score three different MT outputs from the same source text. The QE system can achieve precision and recall above 0.7 simultaneously for MT 2 and MT 3, but not for MT 1, which is a better MT output than MT 2 or MT 3 as indicated by the translation error prevalence level.

Therefore, QE system evaluation must be translation-specific when the PR curve is used. It is important to describe the underlying translation, including its error prevalence rate, when precision and recall are reported for any QE system.

Additional Notes

B. Matthew's correlation coefficient (MCC) analysis

Another common classification-outcome-based method is the MCC analysis; for further details, see Task 3 of WMT26 Automated Translation Quality Evaluation Systems Shared Task and the accompanying overview paper, scheduled for publication in October 2026.

C. Evaluate statistical significance

The evaluation of QE systems is performed on a sample of your data. Therefore, it is important to perform statistical significance analysis to determine whether the patterns you observe in the sample data are the result of a real effect or just random chance.

Statistical significance helps you separate meaningful insights from random noise and avoid making misguided decisions. We refer the user to additional resources on how to perform statistical significance analysis for QE system evaluation using classification-outcome-based methods or correlation-based methods.

D. QE score distribution

Another important consideration is QE score distribution. A good QE system must be able to produce scores with enough spread to support a meaningful threshold. If QE scores are too concentrated, especially near the top, small threshold changes can result in large and unstable change, for example, in indicating which segments should be prioritized for review. This ceiling effect makes pass/fail decisions fragile.

A QE system used on already-good human or machine translations may naturally show compressed scores. Therefore, as the human or machine translation quality improves over time, QE systems must evolve in order to be able to effectively separate an increasingly small number of harder-to-detect translation errors from the rest of the translation.

Appendix 3
Special Cases[4]

In real-life production, users sometimes encounter data patterns that look quite different from the typical ones illustrated in the examples above. Such special cases deserve our attention, and some of them are described below.

A. Poor-quality translation or over editing

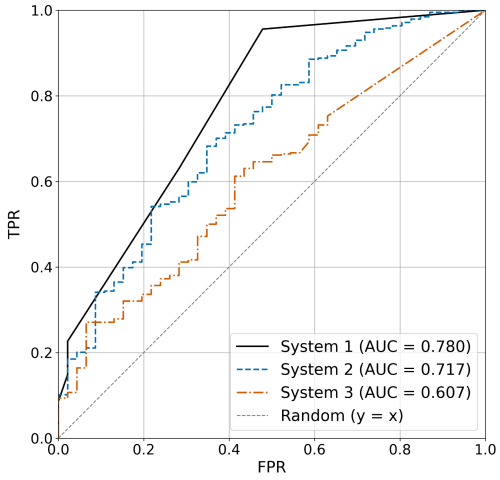


Fig. A1: ROC curves

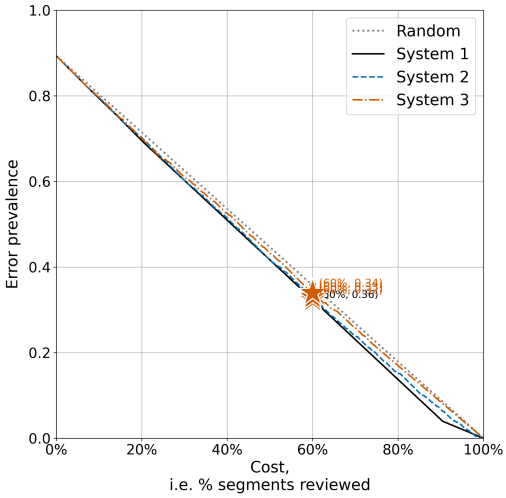


Fig. A2: Error prevalence - % segments reviewed

The ROC curves (Fig. A1 - corresponding to Fig. 1 above) look normal, but when error prevalence is plotted against the % segments reviewed (Fig. A2 - corresponding to Figs. 4 & 7 above), it reveals that none of the QE systems significantly reduces the review volume to achieve a given target error rate as compared to a random review. Fig. A2 also reveals that the error rate prior to review is quite high (~90% of segments were edited).

This suggests that either the MT contains many errors, or the human editor over-edited the MT output. To identify the real cause in this special case, the user could either have a second human to review the MT and post-edited translation, or run the same QE systems on similar content post-edited by other human reviewers.

In cases where the MT output quality is extremely poor, QE systems are generally not able to help the user effectively reduce review cost or effort by detecting segments that are most likely to contain errors, because almost all of the segments contain errors and need to be reviewed and corrected anyway.

In cases where over-editing is a common practice, the user needs to be cautious when selecting their gold standard - actual editing data from production in such cases likely do not accurately reflect the true need of editing and is therefore quite noisy for QE system evaluation purposes.

[4] The special cases are contributed by Kris Shang, Lloyd Wang, Grace Wu, and Sanny Yu of the Acta-Transphere team, whose support is greatly acknowledged.

B. High-quality human translation

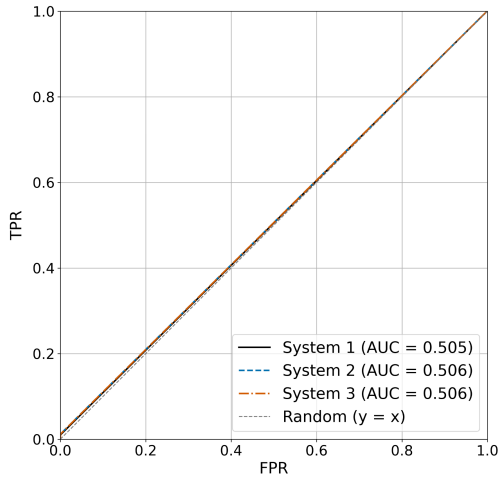


Fig. B1: ROC curves

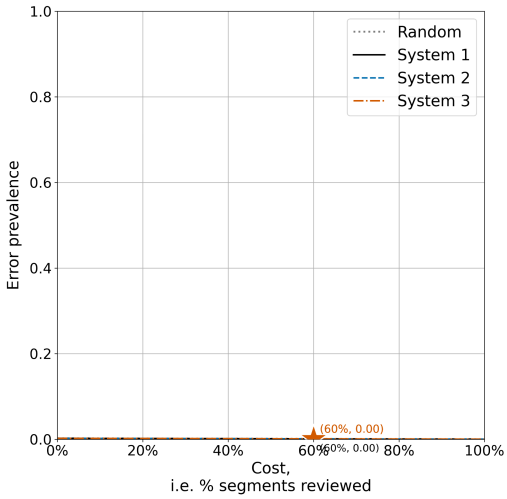


Fig. B2: # of uncorrected error-containing segments - % segments reviewed

The ROC curves of the QE systems are almost identical to that of a random classifier (Fig. B1 - corresponding to Fig. 1 above). The initial error rate in Fig. B2 (corresponding to Figs. 4 & 7 above) reveals why: this translation was almost flawless to begin with. Further investigation confirmed that the translation was a near-perfect human translation.

In such cases where the translation is already very high in quality, it is often challenging for QE systems to detect an extremely low number of minor errors.

C. Low granularity in ROC curves

The user may encounter ROC curves that look less granular, as illustrated in Fig. C1 (corresponding to Fig. 1 above). There are two possible causes: 1. The number of error-containing segments, error-free segments, or both is very small. 2. The QE system outputs a small set of different scores.

In the case illustrated in Fig. C1, an old MT engine was used to produce the translation, resulting in very low quality translation that contained only a small number of error-free segments. Additionally, QE System 1 outputted only a handful of different numeric values for the QE scores, further reducing the granularity of the ROC curve.

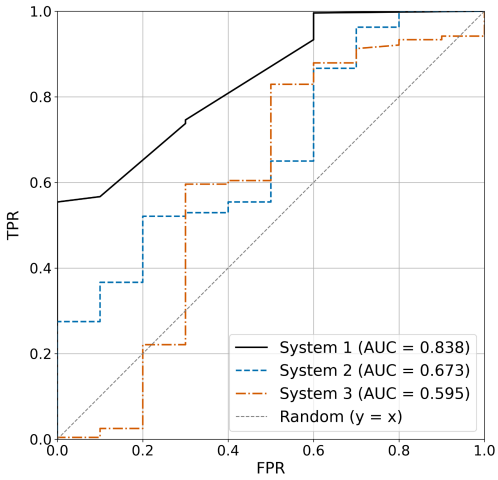


Fig. C1: ROC curves

Appendix 4
Glossary

Term	Definition
AUC	Area Under the Curve: a numerical summary of the total area beneath a plotted curve. In the context of Figure 1 of the Recipe, where ROC curves are plotted, AUC is a summary measure of how well the model separates positives from negatives across all thresholds. In general, higher AUC means better discrimination.
Confusion matrix	A table that compares the model's predicted labels with the gold standard labels. It counts correct and incorrect QE labels across positive and negative classes.
Error	In the context of the Principles and Recipe, "error" refers to a translation error.
Error prevalence	The number of error-containing segments divided by the total number of segments, i.e. the proportion of error-containing segments among all segments.
False negative (FN)	The model predicts negative, but the gold standard is positive. For the examples in Appendix 2, false negative means error-containing segments that are mistakenly predicted as error-free by the QE system.
False negative rate (FNR)	The proportion of actual positives incorrectly labeled negative. Formula: FN / (FN + TP), i.e. FN / P.
False positive (FP)	The model predicts positive, but the gold standard is negative. For the examples in Appendix 2, false positive means error-free segments that are mistakenly predicted as error-containing by the QE system.

Term	Definition
False positive rate (FPR)	The proportion of actual negatives incorrectly labeled positive. Formula: FP / (FP + TN). Since FP + TN = N, the FPR formula can also be expressed as FP / N.
Gold standard	The best available benchmark for your task.
Input translation	The translation that is subject to QE systems for quality estimation. It can be a human translation, human post-edited machine translation, or machine translation.
LLM-as-a-judge QE system	An LLM-as-a-judge QE system is a reference-free translation evaluation system that prompts a large language model (LLM) to assess a machine translation against its source and output predicted quality scores, error spans, error types, error severities, or any combination thereof.
Machine translation (MT)	Non-human translation using computer algorithms.
Negative class (N)	The opposite class from the positive class.
Positive class (P)	The class/event of interest. For the examples in Appendix 2, this means error-containing segment because the interest is in identifying translation errors. However, the positive class can be defined differently for other scenarios. For example, if the interest is in identifying high-quality translation segments, the positive class may be defined as error-free segments.
Precision	Among all items predicted positive, the proportion that is truly positive. Formula: TP / (TP + FP).

Term	Definition
Quality estimation (QE) system	A model or software tool that automatically evaluates translation quality using only the source text and its translation, without relying on a reference translation. A QE system may output: 1. location of errors (error span); 2. estimates of post-editing effort; 3. error-type classification; 4. error-severity ratings; 5. quality scores; or 6. any combination of 1–5.
Recall	Same as TPR: the proportion of actual positives correctly identified.
ROC	Receiver Operating Characteristic curve: a curve showing the trade-off between TPR and FPR across different decision thresholds.
Statistical significance	A measure of whether an observed difference or result is unlikely to have occurred by random chance alone, usually assessed with a p-value or confidence interval.
True negative (TN)	The model predicts negative, and the gold standard is negative. For the examples in Appendix 2, true negative means error-free segments that are correctly predicted as such by the QE system.
True negative rate (TNR)	The proportion of actual negatives correctly identified. Formula: TN / (TN + FP), i.e. TN / N. Also known as specificity.

Term	Definition
True positive (TP)	The model predicts positive, and the gold standard label is positive. For the examples in Appendix 2, true positive means error-containing segments that are correctly predicted as such by the QE system.
True positive rate (TPR)	The proportion of actual positives correctly identified. Formula: TP / (TP + FN). Since TP + FN = P, the TPR formula can also be expressed as TP / P. Also known as sensitivity or recall.