Embedding Similarity Is Not Quality Estimation: Lessons from Replacing a Dedicated QE Model

Dimitrios Zaikis
Welocalize
dimitrios.zaikis@welocalize.com

Andrea Biondo
Welocalize
andrea.biondo@welocalize.com

Matthew Dixon
Welocalize
matthew.dixon@welocalize.com

Konstantinos Karageorgos
Welocalize
konstantinos.karageo@welocalize.com

Aaron Schliem
Welocalize
aaron.schliem@welocalize.com

Abstract

Machine translation quality estimation (QE) typically relies on dedicated neural models trained on human judgments. We evaluate whether cosine similarity over general-purpose embeddings can serve as a lightweight alternative, using Gemini embeddings as the scoring backbone. Through three experiments (rogue dimension analysis, score calibration, and a learned calibration head) and a root cause analysis, we find that cosine similarity between source and translation saturates in the 0.94–0.99 range because even poor translations preserve most of the source semantics, leaving an Area Under the ROC Curve (AUC) ceiling of approximately 0.63. However, a LightGBM classifier trained on normalized cosine and surface-level text features breaks through this ceiling (AUC 0.751), with the improvement driven primarily by features orthogonal to embedding similarity. These findings demonstrate that raw embedding similarity cannot serve as a drop-in QE replacement and identify learned calibration as a viable lightweight path forward.

1 Introduction

Segment-level quality estimation (QE) for machine translation aims to predict whether a translated segment is suitable for release without human review (Zhao et al., 2024). In practice, QE scoring determines how many segments can be approved automatically, directly affecting the cost and speed of the review process (Gladkoff et al., 2024; Cabeça et al., 2023).

The system under study previously used a dedicated neural QE model from the COMET family (Rei et al., 2020), built on XLM-RoBERTa-large and trained on human quality annotations. The replacement was motivated by operational considerations: general-purpose embedding models reduce infrastructure complexity, provide broad multilingual coverage, and were already deployed in the pipeline for other tasks. This model was replaced by a scorer based on cosine similarity over Gemini embeddings (Lee et al., 2025), a general-purpose embedding model that supports a `SEMANTIC_SIMILARITY` task type designed for pairwise text comparison. However, translations inherently preserve most of the source semantics, which may make cosine similarity a weak discriminator of quality differences. The question motivating this work is whether such a general-purpose approach can match the performance of a dedicated QE model, or whether the semantic overlap between source and translation creates a fundamental ceiling.

Initial evaluation revealed a performance difference: AUC 0.631 for the dedicated model versus 0.627 for the embedding scorer. We investigate the causes of this difference and whether calibration or feature engineering can reduce it.

We report on three experiments and a root cause analysis:

1. **Rogue dimension analysis**: Outlier embedding dimensions exist but are harmless; removing them does not change quality rankings (§5.1).

2. **Score calibration**: Redistributing or recalibrating scores cannot improve discrimination

because the ranking is preserved (§5.2).

3. **Learned calibration head**: A LightGBM classifier combining normalized cosine with surface-level features substantially exceeds both baselines (§5.3).

4. **Root cause identification**: Cosine similarity saturates for source-translation pairs because even poor translations preserve semantic content, creating a hard ceiling on discrimination (§5.4).

These experiments explain *why* embedding similarity falls short as a QE proxy and demonstrate that the saturation ceiling can be overcome with a learned model that exploits signals beyond cosine similarity. These findings are directly relevant to localization teams evaluating whether general-purpose embeddings can replace dedicated QE models in production pipelines (Sun and Wang, 2024; Zhao et al., 2024).

2 Background

Quality Estimation. QE aims to predict translation quality without reference translations (Zhao et al., 2024). State-of-the-art QE models such as COMET (Rei et al., 2020), CometKiwi (Rei et al., 2022), and xCOMET (Guerreiro et al., 2024) use cross-lingual encoders (e.g., XLM-RoBERTa) to produce source and translation embeddings, then construct a feature vector $[\mathbf{e}_s; \mathbf{e}_t; |\mathbf{e}_s - \mathbf{e}_t|; \mathbf{e}_s \odot \mathbf{e}_t]$, where $\mathbf{e}_s$ and $\mathbf{e}_t$ are the source and translation embeddings, ; denotes concatenation, and $\odot$ the element-wise product. This vector feeds into a trained regression head. The design explicitly preserves element-wise differences, which allows the model to detect subtle quality signals even when overall semantic similarity is high.

Embedding Similarity for Evaluation. Recent work has explored embedding similarity as a lightweight alternative to trained QE models. The underlying premise is that a faithful translation preserves the semantic content of the source, so cosine similarity between their embeddings should correlate with translation quality: higher similarity implies better preservation of meaning. This reduces QE to a semantic similarity measurement, avoiding the need for quality-annotated training data. Sun et al. (2024) investigate textual similarity as a metric for MT quality estimation with mixed results depending on task formulation. Dinh et al. (2024) propose a k-nearest neighbors approach over embedding representations for QE. Steck et al. (2024) show that cosine similarity over embeddings does not always capture the properties users expect.

Rogue Dimensions. Timkey and van Schijndel (2021) show that transformer language models contain a small number of "rogue dimensions" with abnormally large means that can dominate cosine similarity. They propose z-score standardization as a fix. Whether this affects modern embedding models used for QE has not, to our knowledge, been investigated.

3 System Description

The QE system under study operates as a component within a translation pipeline. For each segment, it produces a quality score used to make a binary accept/reject decision via a per-group threshold.[1]

Dedicated QE Model. The dedicated model uses XLM-RoBERTa-large (1024-dimensional embeddings) to encode source and translation independently, constructs a 4096-dimensional feature vector via concatenation and element-wise operations, and passes this through a trained regression head. The raw model output is rescaled to $[0, 1]$ through post-processing normalizations.

Embedding-Based Scorer. The replacement scorer uses Gemini embeddings (`gemini-embedding-001`) with the `SEMANTIC_SIMILARITY` task type (Lee et al., 2025). Source and translation are embedded independently into 1536-dimensional vectors.[2] The scoring steps are:

1. Compute cosine similarity: $\cos(\mathbf{e}_s, \mathbf{e}_t)$
2. Rescale to $[0, 1]$: $(\cos + 1)/2$
3. Apply post-processing normalizations
4. Compare against the per-group threshold for the accept/reject decision

[1] A group is defined by account, language pair, and content domain.

[2] Truncated from the model's maximum 3072 dimensions via its built-in variable-dimensionality support.

Account	N	% Perf.	Src	Top targets
A	4,318	38.1	en, sv	zh, pt, ru, fr
B	2,666	36.2	en	de, ja, fr, es
C	1,032	56.0	en	fr, es, de, it
D	700	27.4	en	ja, es, fr, zh
E	687	39.0	en	pt, it, ja, zh
F	597	78.2	en	fr, es, pt, it
Total	**10,000**	**41.1**		**38 target langs**

Table 1: Dataset summary. Accounts are anonymized. Source languages are predominantly English with minor Swedish source. "% Perf." is the proportion labeled as acceptable.

4 Experimental Setup

Dataset. We sampled 10,000 translation segments from a pool of 2.1 million segments across six accounts (Table 1). The data are proprietary production content from a large language service provider; no raw translation text is disclosed. The segments are predominantly English source text translated into 38 target languages, covering domains such as technical documentation, user interfaces, and support content. Each segment carries an independent human quality annotation indicating whether the translation meets quality standards: 41.1% are labeled as acceptable, 58.9% require review.

Evaluation Metrics. We evaluate using three metrics that reflect both statistical discrimination and practical applicability:

- **AUC-ROC**: threshold-independent measure of the scorer's ability to separate acceptable from non-acceptable segments.
- **Qualifying accounts**: number of accounts (out of 6) where a threshold can be found satisfying minimum precision (≥ 0.8), minimum support (≥ 50 segments), and minimum volume constraints.
- **Acceptance rate**: percentage of words accepted at the optimal threshold, a key throughput indicator.

Protocol. We use a 70/30 train/test split stratified by the quality label (7,000/3,000 segments). All calibration models are fit on the training set and evaluated on the held-out test set. All data were used in accordance with applicable data processing agreements.

Finding	Value	Interpretation
Top rogue dim (115)	μ=-0.242	$11\times$ std
Top-5 contribution	24.9%	$\approx$1/4 of total score
r^2 after removal	0.9991	Near-perfect rank preservation
Max per-account Δ	0.14%	No group $> 5\%$

Table 2: Rogue dimension analysis. Despite accounting for nearly a quarter of total cosine similarity, rogue dimensions act as constant offsets: removing them changes rankings by only 0.09%.

5 Experiments and Results

5.1 Experiment 1: Rogue Dimension Analysis

Hypothesis. Following Timkey and van Schijndel (2021), Gemini embeddings may contain rogue dimensions whose abnormally large means dominate cosine similarity and distort quality rankings. Removing them might improve discrimination.

Method. We compute per-dimension mean and standard deviation across the full corpus of 20,000 embedding vectors (10K source + 10K target). We flag dimensions where $|\mu_d|$ exceeds 10 standard deviations of the distribution of dimension means. We then measure: (a) how much the top-k rogue dimensions contribute to total cosine similarity, and (b) the r^2 between original and post-removal scores to see whether rankings change.

Results. Rogue dimensions are present: dimension 115 has a mean of -0.242, roughly $11\times$ the standard deviation of all dimension means, and the top 5 dimensions account for 24.9% of total cosine similarity. However, as Table 2 shows, removing them has negligible impact on rankings.

The explanation is that rogue dimensions add a near-constant term to all similarity scores. Since the accept/reject decision depends on *rankings* (threshold-based classification), a constant offset is harmless. This extends the findings of Timkey and van Schijndel (2021) to Gemini embeddings: rogue dimensions exist but do not distort quality rankings.

5.2 Experiment 2: Score Calibration

Hypothesis. The raw cosine similarity scores may be poorly calibrated for the accept/reject decision. Post-processing could improve discrimination.

Methods. Three calibration approaches of increasing complexity were tested:

Method	AUC	Qual.	Acc. %
Dedicated model (COMET)	*0.631*	—	—
Baseline (cosine)	0.627	1/6	14.4
Monotonic norm.	0.627	1/6	14.4
Isotonic regression	0.627	1/6	17.1
Feature combination	0.561	0/6	0.0
LightGBM cal. head	0.751	2/6	34.8

Table 3: Score calibration results on the test set (N=3,000). The dedicated COMET model is shown for reference (threshold-based metrics not directly comparable due to different score distributions). "Qual." = accounts meeting minimum quality criteria. "Acc. %" = best per-account word acceptance rate at optimal threshold. The LightGBM calibration head (§5.3) is trained on normalized cosine and surface-level features.

1. **Monotonic normalization**: an unsupervised monotonic transform that maps scores to a uniform distribution on $[0, 1]$.

2. **Isotonic Regression** (IR): a supervised, non-parametric method that learns a monotone mapping from raw scores to empirical probabilities of being acceptable (Niculescu-Mizil and Caruana, 2005).

3. **Feature Combination** (FC): logistic regression over multiple features extracted from the embedding pairs, including cosine similarity and various distance measures.

Results. Table 3 presents the results. Both the monotonic normalization and isotonic regression produce AUC identical to the baseline (0.627). This confirms a mathematical property of AUC: monotonic transforms preserve rankings by definition and therefore cannot change it.

The feature combination approach performs *worse* than the baseline (AUC 0.561 vs. 0.627), with zero qualifying accounts. Gemini embeddings are perfectly L2-normalized by design, so the norm ratio across all 10,000 pairs has zero variance, making norm-based features constant and uninformative. The remaining distance-based features are monotonic functions of cosine similarity for L2-normalized vectors, so they carry no additional signal.

5.3 Experiment 3: Learned Calibration Head

Hypothesis. A learned model that combines embedding-derived features with surface-level text quality indicators may overcome the cosine saturation ceiling by exploiting signals orthogonal to semantic similarity.

Method. We train a LightGBM (Ke et al., 2017) gradient-boosted tree classifier to predict whether a segment is acceptable, framing QE as binary classification. The feature set combines embedding-derived similarity measures with surface-level text quality indicators. Training uses binary cross-entropy loss with word-count-weighted samples, aligning the objective with the downstream throughput metric (accepted words rather than accepted segments). Hyperparameters are tuned via cross-validation on the training set.

Crucially, the model does *not* use the source and target embeddings directly (no concatenation, no element-wise operations). All embedding information is mediated through cosine similarity and its normalized variants. This design reflects the finding that the 1536-dimensional embeddings contain mostly linguistic content rather than quality signal, making high-dimensional features noisy for a 7,000-sample dataset.

Results. The LightGBM calibration head achieves AUC 0.751 (Table 3), substantially above both the cosine baseline (0.627) and the dedicated COMET model (0.631). Two of six accounts qualify for deployment (vs. one for the cosine baseline), and the best per-account acceptance rate reaches 34.8% (vs. 14.4%). In operational terms, this means substantially more translation volume can be auto-approved, reducing human review load.

Feature importance analysis reveals that surface-level features collectively outweigh the embedding-derived features. This confirms that the improvement comes primarily from signals orthogonal to embedding similarity, capturing error patterns that cosine similarity cannot detect.

5.4 Root Cause: Score Saturation

The experiments above show that the AUC ceiling of around 0.63 is not caused by rogue dimensions, poor score calibration, or unexploited embedding features. The root cause is simpler: **cosine similarity measures semantic overlap, and translations inherently have near-perfect semantic overlap with their source.**

Even a poor translation typically covers the same topic, mentions the same entities, and uses related vocabulary, all of which embedding similarity captures. The quality-relevant differences

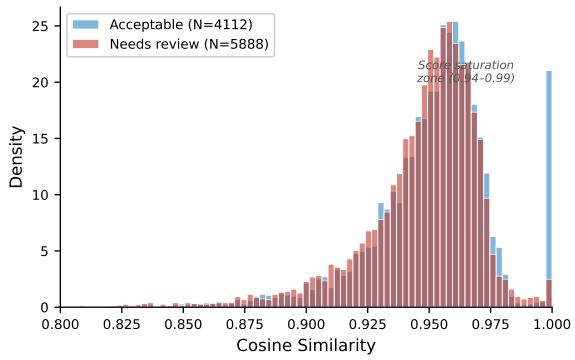

Figure 1: Distribution of cosine similarity scores for acceptable vs. non-acceptable segments. Both distributions overlap heavily in the 0.94–0.99 range, leaving minimal room for threshold-based separation.

(a dropped negation, an incorrect number, wrong terminology) are subtle linguistic details that produce negligible movement in a 1536-dimensional embedding space.

Empirically, cosine similarity scores cluster in the 0.94–0.99 range for both acceptable and non-acceptable segments (Figure 1). The mean score difference between the two classes is only 0.004 (acceptable: 0.951, needs review: 0.947), compressed into a range of roughly 0.05 units, which severely limits threshold-based discrimination.

This contrasts with how dedicated QE models work. The $[\mathbf{e}_s; \mathbf{e}_t; |\mathbf{e}_s - \mathbf{e}_t|; \mathbf{e}_s \odot \mathbf{e}_t]$ feature vector explicitly preserves element-wise differences, and the regression head learns to amplify divergences that signal quality problems. Raw cosine similarity collapses all pairwise information into a single scalar, discarding the fine-grained signal that QE requires. Even the dedicated COMET model only reaches AUC 0.631 on this data (Table 3), which suggests that segment-level QE on heterogeneous multilingual data is inherently challenging.

6 Discussion

Generalizability. We argue that the saturation effect is structural: it follows from what cosine similarity measures, not from the specific embedding model. Any model optimized for semantic similarity will produce high scores for source-translation pairs, because adequate translations *are* semantically similar by definition. The LightGBM result reinforces this: the improvement comes primarily from non-embedding features, confirming that cosine similarity itself provides limited quality signal. Embedding similarity may still be useful for *system-level* comparisons (ranking different MT engines), where outputs genuinely differ in semantic fidelity. The limitation is specific to *segment-level* QE within a single system.

Calibration vs. Discrimination. Monotonic transforms preserve rankings by construction and therefore cannot improve AUC (Niculescu-Mizil and Caruana, 2005). It is important to distinguish calibration problems (fixable via score remapping) from discrimination problems (which require a different model).

Practical Recommendations. Embedding cosine similarity should not be treated as a drop-in replacement for trained QE models. The LightGBM calibration head demonstrates a viable middle ground: a lightweight learned model that combines normalized cosine with surface-level text features can substantially outperform both raw cosine and the dedicated COMET model, without requiring the large-scale human annotation data that dedicated QE models depend on. When a dedicated model is unavailable, such learned calibration approaches are preferable to attempting to post-process similarity scores.

Limitations. This study uses a single embedding model (Gemini), a single pipeline, binary quality labels (not continuous MQM scores), and data from a single domain. Nevertheless, we expect other general-purpose embedding models to exhibit similar saturation, since the effect follows from what cosine similarity measures rather than from the specific model. The AUC difference between models (0.004) is small and we do not report bootstrap confidence intervals; however, the core finding rests not on the magnitude of this difference but on the saturation mechanism that bounds it. Evaluating this hypothesis across other embedding models, and testing the learned calibration approach on larger datasets with continuous quality labels, are immediate next steps.

7 Conclusion

This paper presented a systematic investigation into using Gemini embedding cosine similarity as a replacement for dedicated neural QE. Rogue dimension analysis and score calibration experiments confirm that cosine similarity saturates for source-translation pairs because even poor translations preserve most of the source semantics, resulting in an AUC ceiling of approximately 0.63.

This ceiling cannot be addressed through calibration or feature engineering; it reflects a fundamental limitation of reducing a high-dimensional pairwise comparison to a single similarity scalar.

However, the saturation ceiling is not a hard limit on the data itself. A LightGBM calibration head trained on normalized cosine and surface-level text features achieves AUC 0.751, substantially exceeding both the cosine baseline and the dedicated COMET model. The improvement is driven primarily by features orthogonal to embedding similarity, confirming that the quality signal cosine collapses can be recovered through complementary indicators. This suggests that the path forward for lightweight QE lies in learned models that combine embedding-derived signals with surface-level text features, rather than in better embeddings or more sophisticated similarity measures.

References

Cabeça, Mariana, Marianna Buchicchio, Madalena Gonçalves, Christine Maroti, João Godinho, Pedro Coelho, Helena Moniz, and Alon Lavie. 2023. Quality fit for purpose: Building business critical errors test suites. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 451–460, Tampere, Finland, June. European Association for Machine Translation.

Dinh, Tu Anh, Tobias Palzer, and Jan Niehues. 2024. Quality estimation with k-nearest neighbors and automatic evaluation for model-specific quality estimation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 133–146, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Gladkoff, Serge, Lifeng Han, Gleb Erofeev, Irina Sorokina, and Goran Nenadic. 2024. MTUncertainty: Assessing the need for post-editing of machine translation outputs by fine-tuning OpenAI LLMs. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 337–346, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Guerreiro, Nuno M., Ricardo Rei, Daan Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 3149–3157.

Lee, Jinhyuk, Feiyang Chen, Sahil Dua, Daniel Cer, and Madhuri Shanbhogue. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Niculescu-Mizil, Alexandru and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632. ACM.

Rei, Ricardo, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Yu, António C. Guerreiro, José G. C. de Souza, and Alon Lavie. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. Association for Computational Linguistics.

Steck, Harald, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*. ACM.

Sun, Kun and Rong Wang. 2024. Textual similarity as a key metric in machine translation quality estimation. *arXiv preprint arXiv:2406.07440*.

Timkey, William and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4527–4546. Association for Computational Linguistics.

Zhao, Haofei, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. From handcrafted features to LLMs: A brief survey for machine translation quality estimation. *arXiv preprint arXiv:2403.14118*.