MT Quality Evaluation in the Age of LLM-based MT

Kirti Vashee - Tech Evangelist

translated.

Introduction

We live in an era where machine translation (MT) is ubiquitous, and some form of the technology is in production use across thousands of enterprises. A key requirement during the development and deployment period is a measure of translation quality produced by MT engines.

Automated quality measurements, such as BLEU and COMET, are crucial for **assessing MT translation quality** primarily due to the impracticality of consistent human evaluation. While human assessment is ideal, it's slow, expensive, and difficult to scale across numerous MT systems and language combinations. Automated scores provide a cost-effective and rapid approximation of quality, essential for developers needing quick feedback on evolving models.

Though these scores are mere snapshots and can be misleading approximations, they remain necessary for iterative MT development and for providing some objective, albeit imperfect, comparison where extensive human review is unfeasible.

While these metrics are useful for developers during model development and improvement, experts generally agree that these metrics and their scores are less reliable and accurate than evaluations performed by human linguists in a double-blind setting, especially when measuring progress or comparing different MT systems.

The most popular metric currently in use is COMET (Crosslingual Optimized Metric for Evaluation of Translation). The metric was developed to improve MT quality assessment over earlier automatic metrics such as BLEU, METEOR, or TER, which lacked semantic understanding, source-awareness, and were less reliable with language variations.

By focusing on semantics, COMET proved more resilient to changes in word order, synonyms, and paraphrasing, where BLEU might score low. As a result, COMET gained widespread acceptance and use among developers and enterprise buyers during the peak of Neural MT. Although it was acknowledged to be less precise than human evaluations, it served as a sufficiently close indicator for quick, rough quality assessments.

However, as the use of LLM-based MT increases, we hear increasing discussion on the divergence between quality measurement metrics and human evaluations.

The issue of "metric-bias" was a key discussion at WMT24 (a leading MT-related research conference). The WMT24 research found that this process can lead to *reward hacking*, where the system learns to produce outputs that are favored by the metric, even if these outputs are not better from a human perspective. The research also demonstrated that when the same metric is used both for decoding and for evaluation, the improvements in scores are often overestimated, and documented that there is a systematic penalization of human translations.

The value of automated metrics is increasingly questioned as scores diverge from human evaluations. This is why the Translated team now relies more heavily on human assessments for both internal development and enterprise customer deployments.

Why COMET Scores Diverge from Expert Human Judgments in LLM- Based MT

Despite its strengths, COMET faces several challenges when evaluating translations produced by LLM-based MT systems, including:

- LLM-based translations can be **fluent but hallucinated** or unfaithful to the source. Because COMET training data rarely includes blatant hallucinations with low scores, COMET assigns overly generous values to confident but unfaithful translations. Humans severely penalize such errors, causing a large divergence.

- **Different Quality Dimensions:** Human evaluations consider rich quality aspects (fluency, coherence, style, register, cultural appropriateness) that COMET does not explicitly measure. Thus, valid divergences (idiomatic phrasing, context adaptations, etc.) that a human judge might reward will often reduce the COMET score simply because the wording changed. COMET's training objectives (semantic adequacy) ignore these facets, so ratings diverge when the LLM output is stylistically off yet semantically correct.

- COMET still depends on a **single reference translation**, which may not reflect the diverse outputs possible from LLMs. COMET will penalize correct translations that use different synonyms, word order, or style from the reference. In LLM scenarios, this is critical: an LLM might produce a perfectly good paraphrase or a correct answer with the right nuance that the single reference doesn't capture. Metrics like COMET often emphasize *formal* adequacy to the reference, whereas humans care about *pragmatic* fidelity and fluency.

- Metrics (including COMET) focus on semantic overlap but **miss subtle pragmatic and contextual improvements**: This means an LLM output that is more coherent or culturally appropriate might get a poor COMET score if it differs lexically from the reference. Thus, a key cause of divergence is simply that **humans and COMET "conceptualize translation quality differently"**. COMET captures adequacy to the reference, whereas humans reward broader notions of fidelity and fluency.

- **Empirical Correlation Gaps**: In WMT24 (which included LLM systems), COMET-22's system-level correlation with human MQM ratings was about 0.69 – good but not perfect. (Top new metrics like XCOMET achieved ~0.72.) More tellingly, pairwise and segment correlations can be much weaker. COMET's training and reliance on a single reference bias it toward old MT styles; it focuses on adequacy over higher-level fidelity; and it can give misleadingly high scores to odd outputs. **Researchers, therefore, treat COMET scores carefully and continue to augment them with human or fine-grained analyses when evaluating cutting-edge LLM-based translation systems**

- Performance may vary by language pair, domain, or register. LLMs often handle low-resource languages better than earlier systems, where COMET's training may not generalize well.

- **Known COMET Failure Modes**: Recent analysis has revealed concrete cases where COMET behaves counterintuitively. An entirely *empty* translation (zero words) can still score around 0.32. Worse, **even outputs in the *wrong language* sometimes receive scores higher than empty or random outputs in the correct language**. COMET can be "fooled" by pathological cases, which highlights its mismatch with human judgments. Indeed, it's now acknowledged that an "evaluation crisis" looms: automated metrics often disagree with human raters for LLM outputs.

- The Workshop on Machine Translation (WMT) has consistently emphasized the **importance of human assessment alongside automatic metrics** like BLEU and COMET for evaluating MT output. For instance, **the WMT 2024 shared tasks explicitly state that official rankings for participating systems are determined based on human evaluation scores**, even while automatic metrics are also collected and analyzed. This practice reflects a continued understanding that **human judgment remains the gold standard**.

Summary Table of Root Causes for COMET Score Divergence from Human Evaluations

Root Cause	Typical Metric Symptom	Impact on LLM-Based MT
Training data bias (domain, language)	Under/over-scoring OOD content	High
Version/Precision Drift	Non-reproducible scores	Medium
Single Reference Dependence	Penalizes creative paraphrases	High
Sentence-Level Context	Misses discourse errors	Medium
Metric Hacking (MBR)	Inflated automatic gains	High
Hallucination Insensitivity	False sense of adequacy	High
Style & Bias Blindness	Ignores formality, gender errors	Medium

COMET remains valuable for rapid benchmarking, but practitioners must recognize its blind spots, especially when evaluating the highly varied, stylistically rich outputs typical of modern LLM-based MT solutions.

Discussion with COMET

Author Ricardo Rei

To further explore the issues with COMET suitability and measuring translation quality for LLM-based MT in general, I contacted Ricardo Rei, who is one of the original authors of the COMET paper. He graciously agreed to chat about this, even though he is leaving the MT industry, and the following is a summary of the highlights of our conversation.

Overall Summary

While Ricardo believes COMET is effective for segment-level translation (*a point of disagreement*), he concedes it falls short for broader tasks such as document-level translation, measuring contextual nuances accurately, or style guide adherence. We both agree that an LLM-as-a-judge approach is best for nuanced use cases involving context and specific rules. Ricardo suggests complementing LLM-as-a-judge evaluations with other metrics like COMET, MetricX, and CHRF. We agree that no single metric suffices for complex translation scenarios; instead, **multiple quality signals are needed** to better align with competent human evaluations, which remain the most reliable indicators of quality for researchers and enterprise practitioners alike.

Detail of the Conversation

- **COMET's Suitability for LLM-based MT Evaluation:** I initiated a discussion about the suitability of COMET for evaluating LLM-based machine translation, noting recent evidence at Translated with Lara, and within the research community, suggesting COMET does not perform as well with the latest generation of LLMs as it did with NMT. Ricardo felt that COMET still works well for segment-level translation evaluations when comparing different LLMs on test sets, providing reasonably good correlations with human judgment. (*This has not been the experience of the Lara team, who have noticed regular divergence between COMET scores and double-blind human evaluations even at a segment level.)* However, Ricardo emphasized that **COMET's design is limited to segment-by-segment translation quality** and does not account for broader translation tasks such as document-level, layered context-based translation or adherence to style guides

- **LLM as a Judge and Complementary Metrics:** Ricardo suggested that for nuanced use cases involving context, rules, and style guides, an LLM acting as a judge is currently the best approach for evaluation. This LLM judge should assess both if the rules and instructions are followed, and, additionally, focus on translation quality as well. Ricardo also **recommended combining LLM-as-a-judge evaluations with other metrics** such as COMET, MetricX, and CHRF to develop a more comprehensive overall picture. He referenced the paper *Tower Plus*, which demonstrates this combined approach, evaluating both translation quality and compliance with instruction following. Measurement at a document level is still problematic, but he referenced a "Document Context" quality measurement paper using the SLIDE methodology from Microsoft as one possible approach. We both expect that this (document-level quality measurement) will be a fruitful and useful area for future research.

- **Academia-Industry Gap and the Future of MT Evaluation:** Ricardo pointed out a gap between academia and industry, where academia still focuses heavily on traditional segment-level translation with minimal consideration of the increasing use of documents, style guides, or broader contexts in increasingly more nuanced industry use cases. Ricardo suggested that future evaluations explore reward models, which are more general-purpose and can consider the prompts passed to the models. Both of us agreed that **relying on a single metric is insufficient for complex translation scenarios**, and a combination of multiple quality signals is necessary and required to align with human evaluations.

- **Evolution of Machine Translation Complexity:** Ricardo noted that machine translation has become more complex, moving beyond just understandability and meaning retention to include factors like terminology application, tone, and adherence to specific stylistic guidelines. We both concluded that it is overly optimistic to assume a single quality metric can capture this increased complexity. Ricardo stated: "*There is no single silver bullet metric that can provide the right answer.*" I echoed this, emphasizing that multiple quality indicators are needed to understand model performance, especially in contextually rich scenarios, where relying solely on segment-level metrics such as COMET to make development or deployment decisions is misaligned with human judgments and leads to sub-optimal decisions and choices.

Conclusions: Relying solely on COMET for MT quality evaluation may not align with human assessments. Therefore, a combination of metrics (e.g., COMET, MetricX) is recommended. Additionally, leveraging an LLM-as-a-judge can help confirm adherence to rules and instructions, and proper contextual application, extending beyond just translation quality. When automated metrics are inconsistent, human evaluations remain the definitive standard.

Automated metrics are mainly used to speed up evaluations in two situations:

1. During System Development

    Engineers use automated scores to quickly test different inputs and see what produces the best output. However, relying solely on these scores, especially with newer NMT and LLM models, can lead to poor decisions or dead ends. If the score changes don't make sense, it's crucial to add small-scale **human evaluations for validation**. More and more, human evaluators are brought in to ensure that automated scores actually match what humans perceive as quality

2. For Enterprise System Deployment

    Automated metrics are also used to decide which system is best for real-world enterprise use. But for businesses, **stylistic consistency, contextual accuracy, and subject-specific precision** are vital. Comparisons based only on automated scores, especially if they use irrelevant test data, can often be misleading in professional settings.

As machine translation quality improves, offering greater linguistic nuance, the need for effective and skilled evaluation of its output grows. The “best”, most skillful translators are likely to provide the most insightful and useful quality judgments. The responsibility of identifying the “best” translation and providing a clear justification for that choice will likely remain solely with professional translators.

Best practices show that **effective evaluations require close monitoring of automated metric scores alongside small-scale human assessments**. If these don't line up, you'll need more rigorous, large-scale, double-blind human evaluations. This ensures that big decisions about development or deployment aren't based on scores alone.

As businesses rely more on MT for core communication, collaboration, and knowledge sharing functions, evaluating output from these technologies must also evolve to match the complexity of real-world scenarios. For the future, we will need multiple indicators of quality, and competent **human evaluations remain the most trustworthy** way to predict future performance.

The Bottom Line

If it matters to your business success, or it has a high impact on your customer's experience, always include professional human evaluations in your decision-making around the use and deployment of MT.

Failure example

Results are computed using [Unbabel/wmt22-cometkiwi-da]

Example 1

Source: *Break a leg!*

Translation A Rompiti una gamba!
Translation B In bocca al lupo!

COMET Score: Translation A 0.8174, Translation B 0.4316
Professional translator evaluation: Translation B is the correct figure of speech, perfectly idiomatic to Italian speakers. Translation A is wrong, because it is word-for-word. The Italian "rompiti una gamba" found in Translation A directly translates to "break your own leg".

Example 2

Source: *Was acclaimed for the excellent performance.*
Whole document: *Jasmine arrived early. She prepared thoroughly. Was acclaimed for the excellent performance.*

Translation A È stato acclamato per l'eccellente performance.
Translation B È stata acclamata per l'eccellente performance.

COMET Score: Translation A 0.8706, Translation B 0.8657
Professional translator evaluation: Translation A is wrong because “acclamato” indicates a masculine subject in Italian. While the document does show the feminine subject, the source segment is missing it, and this results in a Gender bias error. Translation B is correct because “acclamata” indicates a feminine subject in Italian, as it was the case.

Example 3

Source: *Was acclaimed for the excellent performance.*
Style guide rule: *“Use Gender-Neutral Terms: Opt for terms that do not indicate gender, such as "staff," "personnel," or "individual" instead of "businessman/businesswoman," "waiter/waitress."*

Translation A Il manager parla bene del loro progetto.
Translation B La direzione parla bene del loro progetto.

COMET Score: Translation A 0.8750, Translation B 0.8447
Professional translator evaluation: *Translation A is wrong because “acclamato” indicates a masculine subject in Italian. While the document does show the feminine subject, the source segment is missing it, and this results in a Gender bias error. Translation B is correct because “acclamata” indicates a feminine subject in Italian, as it was the case.*

Exemple 4

Source: *Breaking the ice at the party was a challenge..*

Translation A Rompere il ghiaccio alla festa sarà una sfida.
Translation B Superare l'imbarazzo iniziale alla festa è stata una sfida.

COMET Score: Translation A 0.8418, Translation B 0.7847
Professional translator evaluation: Translation A has a grammatical error with the verb tense. “Sarà” is in the future, but source had a past verb tense. Translation B, on the other hand, has the correct past tense with “è stata”.

Example 5

Source: *Relax on the soft sands of Shirahama beach.*

Translation A Rilassati sulla morbida sabbia della spiaggia di Shirahama
Translation B Riposati sulle morbide sabbie della spiaggia di Shirahama

COMET Score: Translation A 0.8740, Translation B 0.8770
Professional translator evaluation: Translation A is correct and fluent, while Translation B is wrong. Although the source says “sands”, plural, the same cannot be done in Italian, as “sabbie” wouldn’t make sense. It could even make people think of “sabbie mobili”, which is quicksand**.** Translation B also says “riposati”, which means “rest”, further altering the source meaning.

Notes on the Examples Provided

There may be alternative COMET models that outperform Unbabel/wmt22-cometkiwi-da, particularly those that utilize a reference. We chose to use reference less metrics because incorporating references would make the quality of the reference itself overly influential on the results.

Larger or more recent models available in the literature were not tested and could yield better performance. To draw robust, scientifically valid conclusions, a more in-depth and comprehensive analysis would be required. The examples presented here are selectively chosen to illustrate specific limitations of COMET.

Reference

Are LLMs Breaking MT Metrics?
https://www2.statmt.org/wmt24/pdf/2024.wmt-1.2.pdf#:~:text=XCOMET,20%203%200.686

Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models
https://aclanthology.org/2025.tacl-1.4.pdf#:~:text=Evaluation%20Issues%20Llama2,assessments%20of%20LLM%20translation%20outputs

Pitfalls and Outlooks in Using COMET
https://aclanthology.org/2024.wmt-1.121.pdf#:~:text=Zh%3A%200,random%2C%20but%20fluent%2C%20output

Mitigating Metric Bias in Minimum Bayes Risk Decoding
https://aclanthology.org/2024.wmt-1.109.pdf

Fine-Tuned Machine Translation Metrics Struggle in Unseen Domains
https://assets.amazon.science/6b/72/85118aac4805b6520d6a53699d04/fine-tuned-machine-translation-metrics-struggle-in-unseen-domains.pdf

The SLIDE metric submission to the WMT 2023 metrics task
https://aclanthology.org/2023.wmt-1.68.pdf

The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities https://aclanthology.org/2024.acl-long.336.pdf

WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects
https://arxiv.org/html/2502.12404v1


translated.